Data Assimilation and Inverse Problems

Andrew Stuart † and Armeen Taeb ‡ *

† Department of Computational and Mathematical Sciences
‡ Department of Electrical Engineering
California Institute of Technology
Pasadena, CA 91125

*Email: astuart@caltech.edu, ataeb@caltech.edu

Introduction

Overview of the Notes

These notes are designed with the aim of providing a clear and concise introduction to the subjects of Inverse Problems and Data Assimilation, and their inter-relations, together with citations to some relevant literature in this area.

In its most basic form, inverse problem theory is the study of how to estimate model parameters from data. Often the data provides indirect information about these parameters, corrupted by noise. The theory of inverse problems, however, is much richer than just parameter estimation. For example, the underlying theory can be used to determine the effects of noisy data on the accuracy of the solution; it can be used to determine what kind of observations are needed to accurately determine a parameter; and it can be used to study the uncertainty in a parameter estimate and, relatedly, is useful, for example, in the design of strategies for control or optimization under uncertainty, and for risk analysis. The theory thus has applications in many fields of science and engineering.

To apply the ideas in these notes, the starting point is a mathematical model mapping the unknown parameters to the observations: termed the "forward" or "direct" problem, and often a subject of research in its own right. A good forward model will not only identify how the data is dependent on parameters, but also what sources of noise or model uncertainty are present in the postulated relationship between forward model and data. For example, if the desired forward problem cannot be solved analytically, then the forward model may be approximated by a simulation; in this case discretization may be considered as a source of error. Once a relationship between model parameters, sources of error, and data is clearly defined, the inverse problem of estimating parameters from data can be addressed. The theory of inverse problems can be separated into two cases: (1) the ideal case where data is not corrupted by noise and is derived from a known perfect model; and (2) the practical case where data is incomplete and imprecise. The first case is useful for classifying inverse problems and determining if a given set of observations can, in principle, provide exact solutions; this provides insight into conditions needed for existence, uniqueness, and stability of a solution. The second case is useful for the formulation of practical algorithms to learn about parameters, and uncertainties in their estimates, and will be the focus of these notes.

A model that has the properties: (a) a solution map exists, (b) is unique, and (c) its behavior changes continuously with input (stability) is termed "well-posed". Conversely, a model lacking any of these properties is termed "ill-posed". Ill-posedness is present is many inverse problems, and mitigating it is an extensive part of the subject. Out of the different approaches for formulating an inverse problem, the Bayesian framework naturally offers the ability to assess quality in parameter estimation, and also leads to a form of well-posedness at the level of probability distributions describing the solution.

The goal of the Bayesian framework is to find a probability measure that assigns a probability to each possible solution for a parameter u, given the data y. Bayes formula states that

$$\mathbb{P}(u|y) = \frac{1}{\mathbb{P}(y)}\,\mathbb{P}(y|u)\,\mathbb{P}(u).$$

It enables calculation of the posterior probability on $u|y$, $\mathbb{P}(u|y)$, in terms of the product of the data likelihood $\mathbb{P}(y|u)$ and the prior information on the parameter encoded in $\mathbb{P}(u)$. The likelihood describes the probability of the observed data y, if the input parameter were set to be u; it is determined by the forward model, and the structure of the noise. The normalization parameter $\mathbb{P}(y)$ ensures that $\mathbb{P}(u|y)$ is a probability measure. There are five primary benefits to this framework: (a) it provides a clear theoretical setting in which the forward model choice, noise model and a priori information are explicit; (b) it provides information about the entire solution space for possible input parameter choices; (c) it naturally leads to quantification of uncertainty and risk in parameter estimates; (d) it is generalizable to a wide class of inverse problems, in finite and infinite dimension and comes with a well-posedness theory useful in these contexts; (e) many algorithms to explore $\mathbb{P}(u|y)$ do not require knowledge of the normalization constant $\mathbb{P}(y)$ and so only the likelihood $\mathbb{P}(y|u)$ and the prior $\mathbb{P}(u)$ are needed.

The first half of the notes is dedicated to studying the Bayesian framework for inverse problems. Techniques such as importance sampling and Markov Chain Monte Carlo (MCMC) methods are introduced; these methods have the desirable property that in the limit of an infinite number of samples they reproduce the full posterior distribution. Since it is often computationally intensive to implement these methods, especially in high dimensional problems, approximate techniques such as approximating the posterior by a Dirac or a Gaussian distribution are discussed.

The second half of the notes cover data assimilation. This refers to a particular class of inverse problems in which the unknown parameter is the initial condition of a dynamical system, and in the stochastic dynamics case the subsequent states of the system, and the data comprises partial and noisy observations of that (possibly stochastic) dynamical system. A primary use of data assimilation is in forecasting, where the purpose is to provide better future estimates than can be obtained using either the data or the model alone. All the methods from the first half of the course may be applied directly, but there are other new methods which exploit the Markovian structure to update the state of the system sequentially, rather than to learn about the initial condition. (But of course knowledge of the initial condition may be used to inform the state of the system at later times). We will also demonstrate that methods developed in data assimilation may be employed to study generic inverse problems, by introducing an artificial time to generate a sequence of probability measures interpolating from the prior to the posterior.

Notation

Throughout the notes we use $\mathbb{N}$ to denote the positive integers $\{1, 2, 3, \cdots\}$ and $\mathbb{Z}^+$ to denote the non-negative integers $\mathbb{N} \cup \{0\} = \{0, 1, 2, 3, \cdots\}$. The matrix I_M denotes the identity on $\mathbb{R}^M$. We use $|\cdot|$ to denote the Euclidean norm corresponding to the inner-product $\langle \cdot, \cdot \rangle$. Square matrix A is positive definite (resp. positive semi-definite)

if the quadratic form $\langle u, Au\rangle$ is positive (resp. non-negative) for all $u \neq 0$. By $|\cdot|_B$ we denote the weighted norm defined by $|v|_B^2 = v^* B^{-1} v$. The corresponding weighted Euclidean inner-product is given by $\langle \cdot, \cdot \rangle_B$ and $\langle \cdot, B^{-1} \cdot \rangle$. We use $\otimes$ to denote the outer product between two vectors: $(a \otimes b)c = \langle b, c \rangle a$. We let $B(u, \delta)$ denote the open ball of radius δ at u, in the Euclidean norm.

Throughout we denote by $\mathbb{P}(\cdot), \mathbb{P}(\cdot|\cdot)$ the pdf of a random variable and of a conditional random variable, respectively. We write

$$\mathbb{E}^{\rho}[f] = \int_{\mathbb{R}^N} f(u)\rho(u)du$$

to denote expectation of $f : \mathbb{R}^N \mapsto \mathbb{R}$ with respect to probability measure with probability density function (pdf) ρ on $\mathbb{R}^N$. Our random variables will almost always have density with respect to Lebesgue meausure, but occasional use of Dirac masses will be required; we will use the notationally convenient convention that Dirac mass at point v has "density" $\delta(\cdot - v)$ or $\delta_v(\cdot)$. When random variable u is distributed according to measure with density ρ we will write $u \sim \rho$. We use $\Rightarrow$ to denote weak convergence of probability measures.

Acknowledgements

These notes were developed out of Caltech course ACM 159 in Fall 2017. The notes were created in latex by the students in the class, based on lectures presented by the instructor Andrew Stuart, and on input from the course TA Armeen Taeb. The individuals responsible for the notes listed in alphabetic order are: Blancquart, Paul; Cai, Karena; Chen, Jiajie; Cheng, Richard; Cheng, Rui; Feldstein, Jonathan; Huang, De; Idíni, Benjamin; Kovachki, Nikola; Lee, Marcus; Levy, Gabriel; Li, Liuchi; Muir, Jack; Ren, Cindy; Seylabi, Elnaz, Schäfer, Florian; Singhal, Vipul; Stephenson, Oliver; Song, Yichuan; Su, Yu; Teke, Oguzhan; Williams, Ethan; Wray, Parker; Zhan, Eric; Zhang, Shumao; Xiao, Fangzhou. Furthermore, the following students have added content beyond the class materials: Parker Wray – the Overview, Jiajie Chen – alternative proof of Theorem 1.10 and proof idea for Theorem 14.3, Fangzhou Xiao – numerical simulation of prior, likelihood & posterior, Elnaz Seylabi & Fangzhou Xiao – catching typographical errors in a draft of these notes, Cindy Ren – numerical simulations to enhance understanding of importance sampling in Examples 6.2 and 6.5, Cindy Ren & De Huang – improving the constants in Theorem 6.3 regarding the approximation error of importance sampling, Richard Cheng & Florian Schäfer – illustrations to enhance understanding of the coupling argument used to study convergence of MCMC algorithms by presenting the finite state-space case, and Ethan Williams & Jack Muir – numerical simulations and illustrations of Ensemble Kalman Filter and Extended Kalman Filter. In addition to the students who developed the notes, we would also like to Tapio Helin (Helsinki) who used the notes in his own course and provided very helpful feedback on an early draft.

The work of AS has been funded by the EPSRC (UK), ERC (Europe) and by AFOSR, ARL, NIH, NSF and ONR (USA). This funded research has helped to shape the presentation of the material here and is gratefully acknowledged.

Warning

These are rough notes, far from being perfected. They are likely to contain mathematical errors, incomplete bibliographical information, inconsistencies in notation and typographical errors. We hope that the notes are nonetheless useful. Please contact Armeen Taeb at ataeb@caltech.edu with any feedback from typos, through mathematical errors and bibliographical omissions, to comments on the structural organization of the material.

Contents

1 Bayesian Framework

[2] In this chapter, we introduce the Bayesian approach to inverse problems in finite dimensions. Given a prior distribution characterizing potential solutions u, and given a forward model mapping u to the data y, including the observational noise, the posterior can be derived using Bayes theorem. Having established a formula for the posterior, we investigate the effect the choice in prior has on our solution. We do this by quantifying error in the posterior in the small noise (approaching zero) limit. This provides intuitive understanding concerning the impact of the prior for determined, overdetermined, and underdetermined systems. Several examples will be discussed.

Let $G : \mathbb{R}^N \to \mathbb{R}^J$. We wish to find $u \in \mathbb{R}^N$ from $y \in \mathbb{R}^J$ where

$$y = G(u) + \eta, \tag{1.1}$$

where $\eta \in \mathbb{R}^J$ is the noise. We view $(u, y) \in \mathbb{R}^N \times \mathbb{R}^J$ as a random variable. In this probabilistic perspective, the solution to the problem is to characterize the random variable (r.v.) $u|y \in \mathbb{R}^N$.

Assumption 1.1. *We make the following assumption about the prior u and the noise η.*

- $u \sim \rho_0(u), u \in \mathbb{R}^N$.
- $\eta \sim \pi(\eta), \eta \in \mathbb{R}^J$.
- *u and η are independent, written $u \perp \eta$.*

Here ρ_0 and π describe the (Lebesgue) probability density functions (pdfs) of the variables u and η respectively. Then $\rho_0(u)$ is the prior pdf, $y|u \sim \pi(y - G(u))$ for each fixed $u \in \mathbb{R}^N$ determines the pdf of the likelihood and $u|y$ is a random variable with pdf ρ^y termed the posterior.

1.1 Bayes Theorem

Bayes theorem is a bridge connecting the prior, likelihood and the posterior.

Theorem 1.2 (Bayes theorem)**.** *Let Assumption 1.1 hold and assume that*

$$Z = Z(y) := \int_{\mathbb{R}^N} \pi(y - G(u))\rho_0(u)du > 0.$$

Then $u|y$ is a random variable on $\mathbb{R}^N$ with pdf given by

$$\rho^y(u) = \frac{1}{Z}\pi(y - G(u))\rho_0(u).$$

Proof. Denote by $\mathbb{P}(\cdot), \mathbb{P}(\cdot|\cdot)$ the pdf of random variable or conditional random variable. Using the definition of conditional probability, we have

$$\mathbb{P}(u, y) = \mathbb{P}(u|y)\,\mathbb{P}(y), \text{ if } \mathbb{P}(y) > 0,$$
$$\mathbb{P}(u, y) = \mathbb{P}(y|u)\,\mathbb{P}(u), \text{ if } \mathbb{P}(u) > 0.$$

Note that the marginal pdf on y is given by

$$\mathbb{P}(y) = \int_{\mathbb{R}^N} \mathbb{P}(u, y) du,$$

and similarly for $\mathbb{P}(u)$. Assume $\mathbb{P}(y) > 0$. Then

$$\mathbb{P}(u|y) = \frac{1}{\mathbb{P}(y)} \mathbb{P}(y|u)\, \mathbb{P}(u) = \frac{1}{Z} \pi(y - G(u)) \rho_0(u) \tag{1.2}$$

for both $\mathbb{P}(u) > 0$ and $\mathbb{P}(u) = 0$. Here we remark that

$$\mathbb{P}(y) = Z = \int_{\mathbb{R}^N} Z\, \mathbb{P}(u|y) du = \int_{\mathbb{R}^N} \pi(y - G(u)) \rho_0(u) du > 0. \qquad \square$$

Thus the assumption that $\mathbb{P}(y) > 0$ is justified and the desired result follows from equation (1.2).

1.2 Examples

We consider several examples and investigate the small noise limits of the posterior measure.

Example 1.3. Let $N = J, G(u) := Au, A \in \mathbb{R}^{J \times J}$ be invertible,

$$y = Au + \eta, \quad \eta \sim N(0, \gamma^2 I),$$

where $0 < \gamma \ll 1$. The prior distribution of u is assumed to satisfy

$$\rho_0 \in C(\mathbb{R}^J, \mathbb{R}^+), \quad 0 < \rho_0(u) \leq \rho_{\max} < +\infty, \quad \forall u \in \mathbb{R}^J$$

The posterior can be derived from Bayes theorem and takes the form

$$\rho^y(u) = \frac{1}{Z} \exp\Big(-\frac{1}{2\gamma^2} |y - Au|^2\Big) \rho_0(u),$$

where

$$Z := \int_{\mathbb{R}^N} \exp\Big(-\frac{1}{2\gamma^2} |y - Au|^2\Big) \rho_0(u) du.$$

Now we study the small noise limit of the posterior. Let $u^+ = A^{-1} y$ and $u \in \mathbb{R}^J$ be any point where, for some $\delta \in (0, 2)$,

$$\gamma^{2-\delta} < |y - Au|^2 = |Au^+ - Au|^2 = |u - u^+|^2_{(A^* A)^{-1}}.$$

Using the fact that $0 < \rho_0(v) \leq \rho_{\max},\ \forall v \in \mathbb{R}^J$, we have

$$\rho^y(u^+) = \frac{1}{Z} \rho_0(u^+) := \frac{1}{Z} \rho_0^+ > 0,$$
$$\rho^y(u) \leq \frac{1}{Z} \exp\Big(-\frac{1}{2} \gamma^{-\delta}\Big) \rho_{\max}.$$

It follows that

$$\frac{\rho^y(u^+)}{\rho^y(u)} \geq \exp\left(\frac{1}{2}\gamma^{-\delta}\right)\frac{\rho_0^+}{\rho_{\max}} \to \infty, \text{ as } \gamma \to 0.$$

This shows that the pdf at u^+ is in order of magnitude larger, on the scale $\exp\left(\frac{1}{2}\gamma^{-\delta}\right)$, than at any point outside a small neighbourhood of u^+ which shrinks to zero with γ. Roughly speaking, u^+ is the point which maximizes the posterior pdf as $\gamma \to 0$. Thus the posterior concentrates on the true value of u as the noise in the observation y shrinks to zero. □

Example 1.4. Let $N = 2, J = 1, \rho_0 \in C(\mathbb{R}^2, \mathbb{R}^+), 0 < \rho_0(u) \leq \rho_{\max} < +\infty$, for all $u \in \mathbb{R}^2$ and

$$y = G(u) + \eta = u_1^2 + u_2^2 + \eta, \quad \eta \sim \pi = N(0, \gamma^2), \quad 0 < \gamma \ll 1.$$

Assume that the observation $y > 0$. Using Bayes theorem, we obtain the posterior distribution

$$\rho^y(u) = \frac{1}{Z}\exp\left(-\frac{1}{2\gamma^2}|u_1^2 + u_2^2 - y|^2\right)\rho_0(u).$$

This posterior concentrates on a manifold: the circle $u_1^2 + u_2^2 = y$. Denote by $A^\pm := \{u \in \mathbb{R}^2 : |u_1^2 + u_2^2 - y|^2 \leq \gamma^{2\pm\delta}\}$, for some $\delta \in (0, 2)$, and $\rho_{\min} = \inf_{u\in A^+} \rho_0(u)$. Since the continuous function $\rho_0(u) > 0$, for all $u \in A^+$, and since A^+ is compact, $\rho_{\min} > 0$. Let $u^+ \in A^+, u^- \in (A^-)^c$. Taking the small noise limit yields

$$\frac{\rho^y(u^+)}{\rho^y(u^-)} \geq \exp\left(-\frac{1}{2}\gamma^{\delta} + \frac{1}{2}\gamma^{-\delta}\right)\frac{\rho_{\min}}{\rho_{\max}} \to \infty, \text{ as } \gamma \to 0^+.$$

Therefore, conditional on $y > 0$, the posterior ρ^y concentrates on the circle with radius $\sqrt{y}$ as $\gamma \to 0$. □

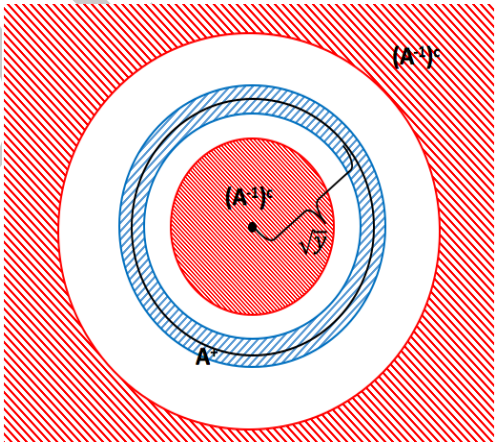


Figure 1 The posterior measure concentrates on a circle with radius $\sqrt{y}$. Here, the blue shadow area is A^+ and the red shadow area is $(A^-)^c$.

Example 1.5. Let $J = N = 1$, $\eta \sim \pi = N(0, \gamma^2)$ and

$$\rho_0(u) = \begin{cases} \frac{1}{2}, & u \in (-1, 1); \\ 0, & u \in (-1, 1)^c. \end{cases}$$

The observation is generated by $y = u + \eta$. Using Bayes Theorem 1.2, we derive the posterior

$$\rho^y(u) = \begin{cases} \frac{1}{2Z} \exp(-\frac{1}{2\gamma^2}|y - u|^2), & u \in (-1, 1); \\ 0, & u \in (-1, 1)^c, \end{cases}$$

where Z is a normalization factor ensuring $\int_{\mathbb{R}} \rho^y(u)du = 1$. The support of ρ^y, i.e. $(-1, 1)$, is the same as the prior ρ_0. Now we find the point which maximizes the posterior pdf. From the explicit formula for ρ^y, we have

$$\arg\max_{u \in \mathbb{R}} \rho^y(u) = \begin{cases} y & \text{if } y \in (-1, 1), \\ -1 & \text{if } y \leq -1, \\ 1 & \text{if } y \geq 1. \end{cases}$$

We remark that anything almost surely true under the prior is also almost surely true under the posterior. In this example, the prior on u is supported on $(-1, 1)$ and the posterior on $u|y$ is supported on $(-1, 1)$. If the data lies in $(-1, 1)$ then the point which is most likely under the posterior is the data itself; otherwise it is the extremal point of the prior support which matches the sign of the data. □

Example 1.6. Let $J = 2, N = 1$, $G : \mathbb{R} \to \mathbb{R}^2$, u be standard Gaussian and

$$y = \begin{pmatrix} 1 \\ 1 \end{pmatrix} u + \eta, \quad \eta \sim N(0, 2\gamma^2 I_2).$$

If $y = (1, -1)^T$, the posterior is given by

$$\begin{aligned} \rho^y(u) &= \frac{1}{Z} \exp\left(-\frac{|y - G(u)|^2}{4\gamma^2}\right) \rho_0(u) \\ &= \frac{1}{Z} \exp\left(-\frac{(1-u)^2}{4\gamma^2} - \frac{(1+u)^2}{4\gamma^2}\right) \cdot \exp\left(-\frac{u^2}{2}\right) \\ &= \frac{1}{Z'} \exp\left(-\frac{(1+\gamma^2)u^2}{2\gamma^2}\right). \end{aligned}$$

Here, we derive the last identity by completing the square and Z' is a new normalization factor. Therefore, $u|y \sim N(0, \frac{\gamma^2}{\gamma^2+1})$. As $\gamma \to 0$, the variance vanishes and the posterior concentrates on 0. However,

$$y = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \neq \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot 0.$$

Thus the model has produced an incorrect inference about which it is very sure. This is caused by model error: the fact that the data is very unlikely to be produced by the forward model in the case $\gamma \ll 1$. □

1.3 Small Noise Limit of the Posterior Distribution: Overdetermined Case

In this section we study the small noise limits in both the over- and underdetermined cases. In particular, we are interested in the limiting behavior of the posterior measure ρ^y as the observational noise tends to zero. For simplicity of the analysis, we assume that the prior distribution is Gaussian. Similar results, however, hold true for other priors.

We assume that $y = G(u) + \eta,\ G : \mathbb{R}^N \to \mathbb{R}^J$ with

$$\eta \sim N(0, B), \quad u \sim \rho_0 = N(m_0, \Sigma_0), \quad G(u) = Au,\ A \in \mathbb{R}^{J\times N}$$

and that B and Σ_0 are both invertible. Then since $y|u \sim N(Au, B)$, the posterior measure ρ^y is a Gaussian $N(m, \Sigma)$. This follows from the fact that the logarithm of ρ^y is quadratic in u under these assumptions. The mean m and variance Σ are given by

$$\begin{aligned} m =& (A^*B^{-1}A + \Sigma_0^{-1})^{-1}(A^*B^{-1}y + \Sigma_0^{-1}m_0) \\ \Sigma =& (A^*B^{-1}A + \Sigma_0^{-1})^{-1}. \end{aligned} \tag{1.3}$$

We will derive these formulae in the next chapter.

Theorem 1.7 (Small Noise Limit of Posterior Distribution- Overdetermined)**.** *Consider the case $N < J$ and assume that $\mathrm{Null}(A) = 0$ and $B = \gamma^2 B_0$. Then in the limit $\gamma^2 \to 0, \rho^y \Rightarrow \delta_{m^+}$, where m^+ is the solution of the least-squares problem and $\Rightarrow$ denotes convergence in distribution.*

$$m^+ = \arg\min_{u\in\mathbb{R}^N} |B_0^{-1/2}(y - Au)|^2. \tag{1.4}$$

Proof. Since $B = \gamma^2 B_0$, we substitute it into (1.3) and deduce

$$\begin{aligned} m =& (A^*B_0^{-1}A + \gamma^2\Sigma_0^{-1})^{-1}(A^*B_0^{-1}y + \gamma^2\Sigma_0^{-1}m_0) \\ \Sigma =& \gamma^2(A^*B_0^{-1}A + \gamma^2\Sigma_0^{-1})^{-1} \end{aligned}$$

Since $\mathrm{Null}(A) = 0$ and B_0 is invertible we deduce that there is $\alpha > 0$ such that

$$\langle \xi, A^*B_0^{-1}A\xi\rangle = |B_0^{-1/2}A\xi|^2 \geq \alpha|\xi|^2, \quad \forall \xi \in \mathbb{R}^N.$$

Thus $A^*B_0^{-1}A$ is positive define and invertible. It follows that as $\gamma \to 0$, the posterior variance $\Sigma \to 0$ and the mean

$$m \to m^* = (A^*B_0^{-1}A)^{-1}A^*B_0^{-1}y.$$

This proves the desired weak convergence of ρ^y to δ_{m^*}. It remains to characterize m^*. Since the null space of A is empty, the minimizers of

$$\varphi(u) := \frac{1}{2}|B_0^{-1/2}(y - Au)|^2$$

are unique and satisfy the normal equations $A^*B_0^{-1}Au = A^*B_0^{-1}y$. Hence m^* solves the desired least-squares problem and thus coincides with m^+ given in (1.4). □

Remark 1.8. In this overdetermined case where A^*B_0A is invertible, the small observational noise limit leads to a posterior which is a Dirac, centered on the solution of a least-square problem determined by the observation operator and the relative weights on the observational noise. Uncertainty disappears, and the prior plays no role in this limit. □

As a byproduct of the above proof of Theorem 1.7, we can determine the limiting behavior of ρ^y in the boundary case $N = J$.

Theorem 1.9 (Small Noise Limit of Posterior Distribution- Determined)**.** *If $N = J$, $\mathrm{Null}(A) = 0$ and $B = \gamma^2 B_0$, then in the limit $\gamma^2 \to 0$, $\rho^y \Rightarrow \delta_{A^{-1}y}$.*

Proof. In the proof of Theorem 1.7, the assumption $N < J$ is used only in that A is not a square matrix and thus A, A^* are not invertible. Denote by (m, Σ) the mean and variance of the posterior $u|y$. Using the same argument, we have $\Sigma \to 0$ and

$$m \to m^* = (A^*B_0^{-1}A)^{-1}A^*B_0^{-1}y$$

Using that A, A^* are square invertible matrices we obtain,

$$m^* = (A^{-1}B_0(A^*)^{-1})A^*B_0^{-1}y = A^{-1}y.$$

Therefore, $\rho^y(u) \Rightarrow \delta^*_m = \delta_{A^{-1}y}$. □

1.4 Small Noise Limit of the Posterior Distribution: Underdetermined Case

In the underdetermined case, we have $N > J$. We assume that $A \in \mathbb{R}^{J\times N}$ with $\mathrm{Rank}(A) = J$ and write

$$A = (A_0\ 0)Q^* = (A_0\ 0)(Q_1\ Q_2)^* = A_0Q_1^* \tag{1.5}$$

with $A_0 \in \mathbb{R}^{J\times J}$ an invertible matrix, $Q = (Q_1\ Q_2) \in \mathbb{R}^{N\times N}$ an orthogonal matrix so that $Q^*Q = I$, $Q_1 \in \mathbb{R}^{N\times J}, Q_2 \in \mathbb{R}^{N\times(N-J)}$. We have the following result:

Theorem 1.10 (Small Noise Limit of Posterior Distribution - Underdetermined)**.** *Let $N > J$ and $B = \gamma^2 B_0$. In the limit $\gamma^2 \to 0$, $\rho^y \Rightarrow N(m^+, \Sigma^+)$, where*

$$\begin{aligned} m^+ &= \Sigma_0 Q_1 (Q_1^*\Sigma_0 Q_1)^{-1} A_0^{-1} y + Q_2 (Q_2^*\Sigma_0^{-1} Q_2)^{-1} Q_2^*\Sigma_0^{-1} m_0 \\ \Sigma^+ &= Q_2 (Q_2^*\Sigma_0^{-1} Q_2)^{-1} Q_2^* \end{aligned}$$

Since $\mathrm{Rank}(\Sigma^+) = rank(Q_2) = N - J < N$ this theorem demonstrates that, in the small observational noise limit, the posterior retains uncertainty in a subspace of dimension $N - J$, and has no uncertainty in a subspace of dimension J.

Example 1.11. To help understand the result in Theorem 1.10, we choose a simple explicit example. Assume that $A = (A_0\ 0) \in \mathbb{R}^{J\times N}, B = \gamma^2 B_0 = \gamma^2 I_J, \Sigma_0 = I_N$, $m_0 = 0$. Let $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim N(0, I_N), u_1 \in \mathbb{R}^J, u_2 \in \mathbb{R}^{N-J}$. The data then satisfies

$$y = Au + \eta = A_0u_1 + \eta,\ \eta \sim N(0, \gamma^2 I_J).$$

The posterior $u|y$ is $\rho_\gamma^y(u) = \frac{1}{Z_\gamma}\exp(-\Phi_\gamma(u;y))$, where Φ_γ is

$$\begin{aligned}\Phi_\gamma(u;y) &= \frac{1}{2\gamma^2}|y - A_0u_1|^2 + \frac{1}{2}|u|^2 \\ &= \left(\frac{1}{2\gamma^2}|y - A_0u_1|^2 + \frac{1}{2}|u_1|^2\right) + \frac{1}{2}|u_2|^2. \end{aligned} \tag{1.6}$$

Consider the contribution to the logarithm of Φ_γ as $\gamma^2 \to 0$. It is clear that $\frac{1}{\gamma^2}|y-A_0u_1|^2$ and completion of the square demonstrates that

$$\rho_\gamma^y(u_1) \Rightarrow \delta_{A_0^{-1}y}(u_1).$$

Once u_1 is fixed as $A_0^{-1}y$, the first term in (1.6) is a constant $\frac{1}{2}|A_0^{-1}y|^2$. Since u_1 and u_2 are independent, formally, we can derive the limiting posterior as follows

$$\rho_\gamma^y(u) \Rightarrow \delta_{A_0^{-1}y}(u_1) \otimes \frac{1}{Z}\exp(-\frac{1}{2}|u_2|^2) = \delta_{A_0^{-1}y}(u_1) \otimes N(0, I_{N-J})$$

where $Z = \int_{\mathbb{R}^{N-J}}\exp(-\frac{1}{2}|u_2|^2)du_2$. In fact, this is exactly the limiting posterior measure given in Theorem 1.10. □

To prove Theorem 1.10, we use the following decomposition of the identity I_N

Lemma 1.12. *Let $\Sigma_0 \in \mathbb{R}^{N\times N}$ be invertible and $Q = (Q_1\ Q_2)$ be an orthonormal matrix with $Q_1 \in \mathbb{R}^{N\times J}, Q_2 \in \mathbb{R}^{N\times(N-J)}$. We have the following decomposition of I_N*

$$I_N = \Sigma_0Q_1(Q_1^*\Sigma_0Q_1)^{-1}Q_1^* + Q_2(Q_2^*\Sigma_0^{-1}Q_2)^{-1}Q_2^*\Sigma_0^{-1} \tag{1.7}$$

Proof. Denote by M the right hand side of (1.7). Since Q is orthonormal, we have $Q_1^*Q_2 = 0, Q_2^*Q_1 = 0$ and thus

$$Q_1^*(M - I) = 0, \quad Q_2^*\Sigma_0^{-1}(M - I) = 0.$$

If $P := (Q_1\ \Sigma_0^{-1}Q_2)$ is full rank, the above identities imply that $P^*(M - I) = 0$ and thus $M = I$. Note that

$$Q^*P = \begin{pmatrix} Q_1^* \\ Q_2^* \end{pmatrix}(Q_1\ \Sigma_0^{-1}Q_2) = \begin{pmatrix} I_J & Q_1^*\Sigma_0^{-1}Q_2 \\ 0 & Q_2^*\Sigma_0^{-1}Q_2 \end{pmatrix}.$$

Since the last matrix is invertible, P is invertible and the proof is complete. □

Proof of Theorem 1.10. Using (1.7), we can decompose u as follows

$$u = \underbrace{\Sigma_0Q_1(Q_1^*\Sigma_0Q_1)^{-1}}_{S}\underbrace{Q_1^*u}_{u_1} + \underbrace{Q_2(Q_2^*\Sigma_0^{-1}Q_2)^{-1}}_{T}\underbrace{Q_2^*\Sigma_0^{-1}u}_{u_2} = Su_1 + Tu_2.$$

Here u_1 and u_2 are Gaussian with $u_2 \sim N(Q_2^*\Sigma_0^{-1}m_0, Q_2^*\Sigma_0^{-1}Q_2)$. The identity

$$\mathrm{Cov}(u_1, u_2) = Q_1^*\mathrm{Cov}(u, u)\Sigma_0^{-1}Q_2 = Q_1^*Q_2 = 0.$$

shows that $u_1 \perp u_2$, that is u_1 and u_2 are independent. From (1.5), we have

$$y = Au + \eta = A_0 Q_1^* u + \eta = A_0 u_1 + \eta. \tag{1.8}$$

Since $u \perp \eta$ and $u_1 \perp u_2$, we know $u_2 \perp y, u_1$. Since the density function of u_1, u_2, η are nonzero everywhere, we apply conditional probability to yield

$$\rho^y(u_1, u_2) := \mathbb{P}(u_1, u_2|y) = \mathbb{P}(u_2)\,\mathbb{P}(u_1|y).$$

Equation (1.8) suggests that (y, u_1, η) with $A_0 \in \mathbb{R}^{J\times J}$ invertible is Gaussian distributed and its posterior is exactly $\mathbb{P}(u_1|y)$. Theorem 1.9 shows that $\mathbb{P}(u_1|y) \Rightarrow \delta_{A_0^{-1}y}(u_1)$ as the noise vanishes, that is as $\gamma^2 \to 0$. Note that $u_2 \perp u_1$ and $u_2 \perp y$. The limiting posterior measure $(u_1, u_2)|y$ is

$$\rho^y(u_1, u_2) \Rightarrow \mathbb{P}(u_2) \otimes \delta_{A_0^{-1}y}(u_1) \tag{1.9}$$

as $\gamma^2 \to 0$. Recall $u = Su_1 + Tu_2$ and $u_2 \sim N(Q_2^*\Sigma_0^{-1}m_0, Q_2^*\Sigma_0^{-1}Q_2)$. The mean and variance of the limiting posterior measure $u|y$ is

$$\begin{aligned} m^+ &= E(Su_1 + Tu_2|y) = SA_0^{-1}y + TE(u_2) = SA_0^{-1}y + TQ_2^*\Sigma_0^{-1}m_0 \\ \Sigma^+ &= Var(Su_1 + Tu_2|y) = Var(Tu_2) = TQ_2^*\Sigma_0^{-1}Q_2T^* = Q_2(Q_2^*\Sigma_0^{-1}Q_2)^{-1}Q_2^* \end{aligned}$$

We have thus completed the proof. □

Remark 1.13. Equation (1.9) shows that in the limit of zero observational noise, the uncertainty is only in the variable u_2. Since $\mathrm{Span}(T) = \mathrm{Span}(Q_2)$ and $u = SA_0^{-1}y + Tu_2$, the uncertainty we observed is in $\mathrm{Span}(Q_2)$. The prior plays a role in the posterior measure, in the limit of zero observational noise, but only in the variables u_2. □

1.5 Discussion and Bibliography

The book by Kaipio and Somersalo [62] provides a good introduction to the Bayesian approach to inverse problems, especially in the context of differential equations. An overview of the subject of Bayesian inverse problems in differential equations, with a perspective informed by the geophysical sciences, is the book by Tarantola [103] (see, especially, chapter 5).

Theorem 1.7 and Theorem 1.10 come from [102]. In that paper the Bayesian approach to regularization is reviewed, developing a function space viewpoint on the subject. A well-posedness theory and some algorithmic approaches which are used when adopting the Bayesian approach to inverse problems are introduced. The function space viewpoint on the subject is developed in more detail in the chapter notes of Dashti and Stuart [24]. An application of this function space methodology, for a large-scale geophysical inverse problem, may be found in [79]. The paper [73] demonstrates the potential for the use of dimension reduction techniques from control theory within statistical inverse problems.

2 The Gaussian Setting

The situation in which the prior and noise models are Gaussian, and the forward map $G(\cdot)$ is linear arises frequently in applications. It is also a setting which is highly amenable to analysis. Results under these assumptions are discussed in this chapter.

2.1 Derivation of Posterior Distribution

Assumption 2.1. *We assume in this chapter that the setting of equation (1.1) applies, that Assumption 1.1 holds and that:*

1. *prior knowledge on $u \in \mathbb{R}^N$: $u \sim \rho_0(u) = N(0, C_0)$, where C_0 is positive definite;*
2. *knowledge on noise $\eta \in \mathbb{R}^J$: $\eta \sim \pi(\eta) = N(0, \Gamma)$, where Γ is positive definite;*
3. *$G : \mathbb{R}^N \to \mathbb{R}^J$ is linear, $G(u) = Au$, where $A \in \mathbb{R}^{J\times N}$.*

Under these assumptions, together with the previous ones, we observe that the likelihood on y given u is a Gaussian,

$$y|u = \pi(y - Au) \sim N(Au, \Gamma). \tag{2.1}$$

We further claim that:

Theorem 2.2 (Posterior is Gaussian)**.** *Under Assumptions 2.1 the posterior distribution is also a Gaussian,*

$$u|y \sim \rho^y(u) := N(m, C). \tag{2.2}$$

The posterior mean m and covariance C are given by the following formulae:

$$m = (C_0^{-1} + A^*\Gamma^{-1}A)^{-1}A^*\Gamma^{-1}y, \tag{2.3}$$

$$C = (C_0^{-1} + A^*\Gamma^{-1}A)^{-1}. \tag{2.4}$$

In fact, by Bayes formula and (2.1), we can write that

$$\begin{aligned}
\rho^y(u) &= \frac{1}{Z}\pi(y - Au)\rho_0(u) \\
&= \frac{1}{Z}\exp\big(-\frac{1}{2}|y - Au|_\Gamma^2\big)\exp\big(-\frac{1}{2}|u|_{C_0}^2\big) \\
&= \frac{1}{Z}\exp\big(-\frac{1}{2}|y - Au|_\Gamma^2 - \frac{1}{2}|u|_{C_0}^2\big) \\
&= \frac{1}{Z}\exp\big(-\mathsf{J}(u)\big),
\end{aligned}$$

with

$$\mathsf{J}(u) = \frac{1}{2}|y - Au|_\Gamma^2 + \frac{1}{2}|u|_{C_0}^2. \tag{2.5}$$

Here Z is the normalization constant with respect to u, ensuring that $\int_{\mathbb{R}^N} \rho^y(u)\mathrm{d}u = 1$.

Proof. (Theorem 2.2) Since $\rho^y(u) = \frac{1}{Z}\exp\big(-\mathsf{J}(u)\big)$ with $\mathsf{J}(u)$, given by (2.5), a quadratic function of u, it follows that the posterior distribution $\rho^y(u)$ is Gaussian. Denoting the mean and variance of $\rho^y(u)$ by m and C, we can write $\mathsf{J}(u)$ in the following form

$$\mathsf{J}(u) = \frac{1}{2}|u-m|_C^2 + \text{const}, \tag{2.6}$$

where the constant is with respect to u. Now matching the coefficients of the quadratic and linear terms in equations (2.5) and (2.6), we get

$$\begin{aligned} C^{-1} &= C_0^{-1} + A^*\Gamma^{-1}A, \\ C^{-1}m &= A^*\Gamma^{-1}y. \end{aligned}$$

Therefore equations (2.3) and (2.4) follow. □

2.2 MAP Estimator

Definition 2.3. The *maximum a posterior (MAP) estimator* of the random variable $u|y$ with (posterior) distribution $\rho^y(u)$, is defined as

$$u^* = \arg\max_{u\in\mathbb{R}^N} \rho^y(u)$$

In our Gaussian case, $\rho^y(u) = \frac{1}{Z}\exp\big(-\mathsf{J}(u)\big)$, thus

$$u^* = \arg\min_{u\in\mathbb{R}^N} \mathsf{J}(u).$$

Furthermore, since $\mathsf{J}(u)$ is a quadratic function of u, with $u = m$ as the axis of symmetry, as equation (2.6) shows, we can easily identify m as the MAP estimator of u:

Theorem 2.4 (Characterize MAP Estimator)**.** *The MAP estimator under Assumptions 2.1, is $u^* = m$ where m is given by equation (2.3).*

Example 2.5. Let $\Gamma = \gamma^2 I$, $C_0 = \sigma^2 I$ and set $\lambda = \frac{\sigma^2}{\gamma^2}$. Then

$$\mathsf{J}_\lambda(u) := \gamma^2 J(u) = \frac{1}{2}|y-Au|^2 + \frac{\lambda}{2}|u|^2.$$

Since m minimizes $\mathsf{J}_\lambda(\cdot)$ it follows that

$$(A^*A + \lambda I)m = A^*y \tag{2.7}$$

Remark 2.6. Example 2.5 provides a link between Bayesian inversion and optimization approaches to inversion: $\mathsf{J}_\lambda(u)$ can be seen as the objective functional in a linear regression model with a regularizer $\frac{\lambda}{2}|u|^2$, as used in ridge regression. The equation (2.7) for m is exactly the normal equation with regularizer in the least square problem. In fact, in the general case, equation (2.3) can be also viewed as a generalized normal equation. This point of view helps us understand the structure of Bayesian regularization, by linking it to the deep understanding of optimization approaches to inverse problems. □

2.3 Posterior Consistency

We assume further that in this section

Assumption 2.7. *In addition to the Assumptions 2.1 we also assume that:*

1 $N = J$*;*

2 $A \in \mathbb{R}^{N\times N}$ *is invertible;*

3 $\eta := \gamma\eta_0$ *where* $\eta_0 \sim N(0, \Gamma_0)$ *so that* $\Gamma = \gamma^2\Gamma_0$*;*

4 $y = Au^\dagger + \gamma\eta_0^\dagger$*, for fixed* $u^\dagger, \eta_0^\dagger \in \mathbb{R}^N$*.*

Then we claim that:

Theorem 2.8 (Posterior Consistency)**.** *Let Assumptions 2.7 hold. Then for any sequence* $M(\gamma) \to \infty$ *as* $\gamma \to 0$*, with* $\mathbb{P}$ *denote probability under the posterior distribution,*

$$\mathbb{P}\{|u - u^\dagger|^2 > M(\gamma)\gamma^2\} \to 0. \tag{2.8}$$

Remark 2.9. For any $\varepsilon > 0$, set $M(\gamma) = \frac{\varepsilon^2}{\gamma^2}$ in Theorem 2.8, to obtain

$$\mathbb{P}\{|u - u^\dagger| > \varepsilon\} \to 0.$$

Thus Theorem 2.8 implies that u converges to $u^\dagger$ in probability.

Since C is symmetric positive semi-definite, we can assume that its eigenvalues are λ_i^2, orthogonal eigenvectors are φ_i, $i = 1, 2, \ldots, N$, such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$, $\langle\varphi_i, \varphi_j\rangle = \delta_{ij}$. Therefore, $\langle\varphi_j, C\varphi_j\rangle = \lambda_j^2\langle\varphi_j, \varphi_j\rangle = \lambda_j^2$, for $j = 1, 2, \ldots, N$. To prove the preceding theorem we will use the following lemma.

Lemma 2.10 (Karhunen−Loève Expansion)**.** *If* ξ *is a random variable in* $\mathbb{R}^N$ *and* $\xi \sim N(0, C)$*, then we can write the Karhunen−Loève decomposition of random variable* $\xi \sim N(0, C)$*:*

$$\xi = \sum_{j=1}^{N} \lambda_j\xi_j\varphi_j, \tag{2.9}$$

where the $\{\xi_j\}$ *are a collection of independent* $N(0, 1)$ *variables.*

Proof. To see this let $\xi_j = \frac{1}{\lambda_j}\langle\varphi_j, \xi\rangle$, $j = 1, 2, \ldots, N$. Then by the properties of Gaussian vectors, and in particular the definition of covariance, it follows that

$$\xi_j \sim N\Big(0, \frac{1}{\lambda_j^2}\langle\varphi_j, C\varphi_j\rangle\Big) = N(0, 1).$$

Furthermore, we compute that, if $i \neq j$,

$$\begin{bmatrix}\xi_i \\ \xi_j\end{bmatrix} = \begin{bmatrix}\frac{1}{\lambda_i}\varphi_i^T \\ \frac{1}{\lambda_j}\varphi_j^T\end{bmatrix}\xi \sim N\Big(0, \begin{bmatrix}\frac{1}{\lambda_i}\varphi_i^T \\ \frac{1}{\lambda_j}\varphi_j^T\end{bmatrix} C \begin{bmatrix}\frac{1}{\lambda_i}\varphi_i & \frac{1}{\lambda_j}\varphi_j\end{bmatrix}\Big) = N\Big(0, \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}\Big).$$

Thus $\mathrm{Cov}(\xi_i, \xi_j) = 0$ and since ξ_i and ξ_j are Gaussian this also implies that they are independent. It remains to note that

$$\begin{aligned}\sum_{j=1}^{N} \lambda_j \xi_j \varphi_j &= \sum_{j=1}^{N} \lambda_j \frac{1}{\lambda_j}\langle \varphi_j, \xi\rangle \varphi_j \\ &= \sum_{j=1}^{N} \langle \varphi_j, \xi\rangle \varphi_j = \xi.\end{aligned}$$

Therefore we have the decomposition in equation (2.9). □

Once we have the decomposition for ξ, we may verify the covariance defining property that

$$\begin{aligned}\mathbb{E}[\xi \otimes \xi] &= \mathbb{E}[\sum_{j=1}^{N}\sum_{k=1}^{N} \lambda_j \lambda_k \xi_j \xi_k \varphi_j \otimes \varphi_k] \\ &= \sum_{j=1}^{N}\sum_{k=1}^{N} \lambda_j \lambda_k \delta_{jk} \varphi_j \otimes \varphi_k \\ &= \sum_{j=1}^{N} \lambda_j^2 \varphi_j \otimes \varphi_j = C.\end{aligned}$$

Furthermore, since $|\xi|^2 = \sum_{j=1}^{N} \lambda_j^2 \xi_j^2$, we can verify the formula

$$\mathbb{E}[|\xi|^2] = Tr(C). \tag{2.10}$$

We may now prove our posterior consistency theorem.

Proof. (Theorem 2.8) We start our proof by estimating $e = m - u^\dagger$. From equation (2.3) we have

$$(A^*\Gamma_0^{-1}A + \gamma^2 C_0^{-1})m = A^*\Gamma_0^{-1}y.$$

Using Assumption 2.7(4) we replace y in the right hand side to obtain

$$\begin{aligned}A^*\Gamma_0^{-1}Am + \gamma^2 C_0^{-1}m &= A^*\Gamma_0^{-1}(Au^\dagger + \gamma\eta_0^\dagger) \\ &= A^*\Gamma_0^{-1}Au^\dagger + \gamma A^*\Gamma_0^{-1}\eta_0^\dagger\end{aligned}$$

Subtracting $A^*\Gamma_0^{-1}Au^\dagger + \gamma^2 C_0^{-1}u^\dagger$ from both sides we get

$$(A^*\Gamma_0^{-1}A + \gamma^2 C_0^{-1})e = \gamma A^*\Gamma_0^{-1}\eta_0^\dagger - \gamma^2 C_0^{-1}u^\dagger.$$

We proceed to analyze by the energy method. Taking the inner product of both sides with e we obtain

$$\langle e, A^*\Gamma_0^{-1}Ae\rangle + \gamma^2\langle e, C_0^{-1}e\rangle = \gamma\langle e, A^*\Gamma_0^{-1}\eta_0^\dagger\rangle - \gamma^2\langle e, C_0^{-1}u^\dagger\rangle,$$

that is,

$$|Ae|^2_{\Gamma_0} + \gamma^2|e|^2_{C_0} = \gamma\langle e, A^*\Gamma_0^{-1}\eta_0^\dagger\rangle - \gamma^2\langle e, C_0^{-1}u^\dagger\rangle. \tag{2.11}$$

Now observe that $|A\cdot|_{\Gamma_0}$ defines a norm on $\mathbb{R}^N$, because A is invertible. Since all norms on $\mathbb{R}^N$ are equivalent, there exists $\alpha = \alpha(A, \Gamma_0) > 0$, such that $|Ae|^2_{\Gamma_0} \geq \alpha|e|^2$. And we always have that $\gamma^2|e|^2_{C_0} \geq 0$. Thus the left hand side of equation (2.11) is no smaller than $\alpha|e|^2$. For the right hand side, we denote

$$K = K(A, \Gamma_0, \eta_0^\dagger, C_0, u^\dagger) = 2\max\Big(|A^*\Gamma_0^{-1}\eta_0^\dagger|, |C_0^{-1}u^\dagger|\Big) \geq 0.$$

Note that K is constant under Assumption 2.7(4). Then, when $\gamma \in (0, 1)$, by the Cauchy-Schwartz inequality,

$$\begin{aligned}\gamma\langle e, A^*\Gamma_0^{-1}\eta_0^\dagger\rangle - \gamma^2\langle e, C_0^{-1}u^\dagger\rangle &\leq \gamma|e||A^*\Gamma_0^{-1}\eta_0^\dagger| + \gamma^2|e||C_0^{-1}u^\dagger| \\ &\leq \frac{\gamma}{2}K|e| + \frac{\gamma^2}{2}K|e| \\ &\leq \frac{\gamma}{2}K|e| + \frac{\gamma}{2}K|e| = \gamma K|e|\end{aligned}$$

Put these results together, we have, when $\gamma \in (0, 1)$,

$$\alpha|e|^2 \leq \gamma K|e|.$$

That is,

$$|e| \leq \frac{K\gamma}{\alpha}. \tag{2.12}$$

Now we turn to estimate the trace of C. In order to manipulate the definition of the induced matrix norm, denote $b = C^{-1}a$, then

$$|C| = \sup_{b\in\mathbb{R}^N} \frac{|Cb|}{|b|} = \sup_{b\in\mathbb{R}^N} \frac{|a|}{|b|}.$$

Since $C = (\frac{1}{\gamma^2}A^*\Gamma_0^{-1}A + C_0^{-1})^{-1}$, we have that

$$(A^*\Gamma_0^{-1}A + \gamma^2 C_0^{-1})a = \gamma^2 b.$$

Again using the energy method we find that

$$|Aa|^2_{\Gamma_0} + \gamma^2|a|^2_{C_0} = \gamma^2\langle a, b\rangle. \tag{2.13}$$

As before the left hand side is bounded below by $\alpha|a|^2$; and by the Cauchy-Schwartz inequality, the right hand side is bounded above by $\gamma^2|a||b|$. So from equation (2.13),

$$\frac{|a|}{|b|} \leq \frac{\gamma^2}{\alpha}.$$

By the arbitrariness of a and b,

$$|C| = \sup_{b\in\mathbb{R}^N} \frac{|a|}{|b|} \leq \frac{\gamma^2}{\alpha} \tag{2.14}$$

Since $C^{-1} = \frac{1}{\gamma^2}A^*\Gamma_0^{-1}A + C_0^{-1}$ is a symmetric positive-definite matrix, so C is also symmetric positive definite. Therefore, its eigenvalues are all positive real numbers and we write them as $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_N^2 > 0$. For symmetric positive definite matrix, we have $\sigma_1^2 = |C| \leq \frac{\gamma^2}{\alpha}$, by the inequality (2.14). So here we have,

$$Tr(C) = \sum_{i=1}^{N} \sigma_i^2 \leq N\sigma_1^2 \leq \frac{N\gamma^2}{\alpha} \tag{2.15}$$

Finally, we put our results together to get the conclusion. Since we already know that $u|y \sim N(m, C)$, we denote by ξ the centered random variable $\xi = u|y - m \sim N(0, C)$. Now let $\mathbb{E}$ denote expectation with respect to the posterior distribution, with y given by Assumption 2.1(4). Then when $\gamma \in (0, 1)$, according to inequality (2.12), (2.15) and Lemma 2.10,

$$\begin{aligned}\mathbb{E}[|u - u^\dagger|^2] &= \mathbb{E}[|m - u^\dagger + \xi|^2] = \mathbb{E}[|m - u^\dagger|^2] + \mathbb{E}[|\xi|^2] \\ &= |m - u^\dagger|^2 + \mathbb{E}[|\xi|^2] \\ &= |e|^2 + Tr(C) \leq L\gamma^2,\end{aligned}$$

where $L = \frac{K^2+\alpha N}{\alpha^2}$ is a constant. The final step follows from the Markov inequality: for any $M(\gamma) \to \infty$ when $\gamma \to 0$,

$$\mathbb{P}\{|u - u^\dagger|^2 > M(\gamma)\gamma^2\} \leq \frac{\mathbb{E}[|u - u^\dagger|^2]}{M(\gamma)\gamma^2} \leq \frac{L}{M(\gamma)} \to 0, \text{ as } \gamma \to 0.$$

Equation (2.8) follows. □

2.4 Discussion and Bibliography

The linear Gaussian setting plays a central role in the study of inverse problems, for several reasons. One is that it allows explicit solutions which can be used to give insight into the subject area more generally. The second is that in the large data limit many Bayesian posteriors are approximately Gaussian. The paper [36], which is in the ;inear Gaussian setting, plays an important role in the history of Bayesian inversion as it was arguably the first to formulate Bayesian inversion in function space.

We have also employed the Gaussian setting to present a basic form of posterior consistency in the Bayesian setting. For a treatment in infinite dimensions see [65, 3, 85].

For the consistency problem in the classical statistical setting, see the books [42, 107]. The book [107] also contains definition and properties of convergence in probability as used here. For the non-statistical approach to inverse problems, and consistency results, see [30] and the references therein.

3 Well-posedness and Approximation

In this chapter we show that the Bayesian formulation of inverse problems leads to a form of well-posedness; this is turn may be used to control errors introduced in the posterior distribution by perturbations of various kinds. In order to discuss these issues we will need to introduce metrics on probability measures, and part of the chapter will be devoted this topic.

3.1 Approximation Problem

Recall the inverse problem setting of finding $u \in \mathbb{R}^N$ from $y \in \mathbb{R}^J$ given by (1.1). The noise $\eta \sim \pi(\cdot)$ and prior $u \sim \rho_0$ are given by Assumption 1.1. The posterior $\rho^y(u)$ on $u|y$ is given by Theorem 1.2. For simplicity we will simply write $\rho^y(u) = \rho(u)$ throughout this chapter.

Our goal here is to consider what happens when computation of $G(u)$ is replaced by $G_\delta(u)$, and consequently $\rho(u)$ is replaced by $\rho_\delta(u)$. Such scenario arises when the true $G(u)$ is not accessible but can be approximated by some computable $G_\delta(u)$. A commonly arising situation in applications, to which the theory contained herein may be generalized, arises when $G(u)$ is an operator acting on an infinite-dimensional space which is approximated, for the purposes of computation, by some finite-dimensional operator $G_\delta(u)$. We seek to prove that, under certain assumptions, the small difference between $G(u)$ and $G_\delta(u)$ (forward error) leads to similarly small difference between $\rho(u)$ and $\rho_\delta(u)$ (inverse error):

Meta Theorem: Well-posedness

$$|G(u) - G_\delta(u)| = O(\delta) \implies d(\rho, \rho_\delta) = O(\delta),$$

for small enough $\delta > 0$ and some metric $d(\cdot, \cdot)$ on probability densities.

We will show that the $O(\delta)$-convergence of ρ_δ with respect to some $d(\cdot, \cdot)$ can be guaranteed under certain assumptions, and we will give an example where these assumptions hold true.

3.2 Metrics on Probability Densities

We first introduce two frequently-used metrics that introduce a measure of distance between probability densities.

- The *total variation distance* between two probability densities ρ, ρ' is defined by

$$d_{\mathtt{TV}}(\rho, \rho') := \frac{1}{2} \int |\rho(x) - \rho'(x)| dx = \frac{1}{2} \|\rho - \rho'\|_{L^1}.$$

- The *Hellinger distance* between two probability densities ρ, ρ' is defined by

$$d_{\mathtt{Hell}}(\rho, \rho') := \Big(\frac{1}{2} \int |\sqrt{\rho(x)} - \sqrt{\rho'(x)}|^2 dx \Big)^{\frac{1}{2}} = \frac{1}{\sqrt{2}} \|\sqrt{\rho} - \sqrt{\rho'}\|_{L^2}.$$

We here prove some properties of these two metrics that may be used in future chapters.

Lemma 3.1. *For any probability densities ρ, ρ',*

$$0 \le d_{TV}(\rho, \rho') \le 1, \quad 0 \le d_{Hell}(\rho, \rho') \le 1.$$

Proof. The lower bounds follow immediately from the definitions. We only need prove the upper bounds:

$$d_{\mathtt{TV}}(\rho, \rho') = \frac{1}{2}\int |\rho(x) - \rho'(x)|dx \le \frac{1}{2}\int \rho(x)dx + \frac{1}{2}\int \rho'(x)dx = 1,$$

$$\begin{aligned} d_{\mathtt{Hell}}(\rho, \rho') &= \Big(\frac{1}{2}\int |\sqrt{\rho(x)} - \sqrt{\rho'(x)}|^2 dx\Big)^{\frac{1}{2}} \\ &= \Big(\frac{1}{2}\int (\rho(x) + \rho'(x) - 2\sqrt{\rho(x)\rho'(x)})dx\Big)^{\frac{1}{2}} \\ &\le \Big(\frac{1}{2}\int (\rho(x) + \rho'(x))dx\Big)^{\frac{1}{2}} \\ &= 1. \end{aligned}$$

□

Lemma 3.2. *For any probability densities ρ, ρ',*

$$\frac{1}{\sqrt{2}} d_{TV}(\rho, \rho') \le d_{Hell}(\rho, \rho') \le \sqrt{d_{TV}(\rho, \rho')}.$$

Proof. We use the Cauchy–Schwartz inequality to prove that

$$\begin{aligned} d_{\mathtt{TV}}(\rho, \rho') &= \frac{1}{2}\int |\sqrt{\rho(x)} - \sqrt{\rho'(x)}||\sqrt{\rho(x)} + \sqrt{\rho'(x)}|dx \\ &\le \Big(\frac{1}{2}\int |\sqrt{\rho(x)} - \sqrt{\rho'(x)}|^2 dx\Big)^{\frac{1}{2}} \Big(\frac{1}{2}\int |\sqrt{\rho(x)} + \sqrt{\rho'(x)}|^2 dx\Big)^{\frac{1}{2}} \\ &\le d_{\mathtt{Hell}}(\rho, \rho')\Big(\frac{1}{2}\int (2\rho(x) + 2\rho'(x))dx\Big)^{\frac{1}{2}} \\ &= \sqrt{2} d_{\mathtt{Hell}}(\rho, \rho'). \end{aligned}$$

Notice that $|\sqrt{\rho(x)} - \sqrt{\rho'(x)}| \le |\sqrt{\rho(x)} + \sqrt{\rho'(x)}|$ since $\sqrt{\rho(x)}, \sqrt{\rho'(x)} \ge 0$. Thus we have

$$\begin{aligned} d_{\mathtt{Hell}}(\rho, \rho') &= \Big(\frac{1}{2}\int |\sqrt{\rho(x)} - \sqrt{\rho'(x)}|^2 dx\Big)^{\frac{1}{2}} \\ &\le \Big(\frac{1}{2}\int |\sqrt{\rho(x)} - \sqrt{\rho'(x)}||\sqrt{\rho(x)} + \sqrt{\rho'(x)}|dx\Big)^{\frac{1}{2}} \\ &\le \Big(\frac{1}{2}\int |\rho(x) - \rho'(x)|dx\Big)^{\frac{1}{2}} \\ &= \sqrt{d_{\mathtt{TV}}(\rho, \rho')}. \end{aligned}$$

□

Lemma 3.3. *Let f be a function such that $\sup_{u\in\mathbb{R}^N}|f(u)| \le f_{\max} < \infty$, then*

$$\left|\mathbb{E}^{\rho}[f] - \mathbb{E}^{\rho'}[f]\right| \le 2f_{\max} d_{TV}(\rho,\rho').$$

Proof.

$$\begin{aligned}\left|\mathbb{E}^{\rho}[f] - \mathbb{E}^{\rho'}[f]\right| &= \Big|\int_{\mathbb{R}^N} f(u)(\rho(u)-\rho'(u))du\Big| \\ &\le 2f_{\max}\cdot\frac{1}{2}\int_{\mathbb{R}^N}|\rho(u)-\rho'(u))|du \\ &= 2f_{\max} d_{\mathtt{TV}}(\rho,\rho').\end{aligned}$$

□

Lemma 3.4. *Let f be a function such that $\mathbb{E}^{\rho}[|f|^2] + \mathbb{E}^{\rho'}[|f|^2] \le F^2 < \infty$, then*

$$\left|\mathbb{E}^{\rho}[f] - \mathbb{E}^{\rho'}[f]\right| \le 2F d_{Hell}(\rho,\rho').$$

Proof.

$$\begin{aligned}\left|\mathbb{E}^{\rho}[f] - \mathbb{E}^{\rho'}[f]\right| &= \Big|\int_{\mathbb{R}^N} f(u)(\sqrt{\rho(u)}-\sqrt{\rho'(u)})(\sqrt{\rho(u)}+\sqrt{\rho'(u)})du\Big| \\ &\le \Big(\frac{1}{2}\int|\sqrt{\rho(u)}-\sqrt{\rho'(u)}|^2du\Big)^{\frac{1}{2}}\Big(2\int|f(u)|^2|\sqrt{\rho(u)}+\sqrt{\rho'(u)}|^2du\Big)^{\frac{1}{2}} \\ &= d_{\mathtt{Hell}}(\rho,\rho')\Big(4\int|f(u)|^2(\rho(u)+\rho'(u))du\Big)^{\frac{1}{2}} \\ &= 2F d_{\mathtt{Hell}}(\rho,\rho').\end{aligned}$$

□

3.3 Main Theorem

We write

$$r(u) = \sqrt{\pi(y-G(u))} \quad \text{and} \quad r_\delta(u) = \sqrt{\pi(y-G_\delta(u))},$$

which then gives

$$\rho(u) = \frac{1}{Z}r(u)^2\rho_0(u) \quad \text{and} \quad \rho(u) = \frac{1}{Z_\delta}r_\delta(u)^2\rho_0(u),$$

where $Z = \int r^2\rho_0(u)du$ and $Z_\delta = \int r_\delta(u)^2\rho_0(u)du$. Before we proceed to our main result, we first make some assumptions:

Assumption 3.5. *$\exists\delta_c > 0$, $K_1, K_2 < \infty$ such that $\forall\delta\in(0,\delta_c)$ we have*

(i) $|r(u) - r_\delta(u)| \le L(u)\delta$, for some $L(u)$ such that $\mathbb{E}^{\rho_0}[L^2(u)] \le K_1$;

(ii) $\sup_{u\in\mathbb{R}^N}(|r(u)|+|r_\delta(u)|)\le K_2$;

(iii) $Z>0$.

Now we state the main theorem of this chapter:

Theorem 3.6 (Well-posedness of Posterior)**.** *Under Assumption 3.5 we have*

$$d_{\mathtt{Hell}}(\rho,\rho_\delta)\le C\delta,\quad \delta\in(0,\tilde\delta_c),$$

for some $\tilde\delta_c>0$ *and some* $C\in(0,+\infty)$ *independent of* δ.

To prove Theorem 3.6, we first prove a lemma which characterizes the normalization factor Z_δ in the small δ limit.

Lemma 3.7. *Under Assumption 3.5* $\exists\tilde\delta_c>0$, $c_1,c_2\in(0,+\infty)$ *such that*

$$|Z-Z_\delta|\le c_1\delta\quad \textit{and}\quad Z,Z_\delta>c_2,\quad \textit{for } \delta\in(0,\tilde\delta_c).$$

Proof. Since $Z=\int r^2(u)\rho_0(u)du$ and $Z_\delta=\int r_\delta^2(u)\rho_0(u)du$ we have

$$\begin{aligned}
|Z-Z_\delta| &= \int (r^2(u)-r_\delta^2(u))\rho_0(u)du\\
&\le \Big(\int |r(u)-r_\delta(u)|^2\rho_0(u)du\Big)^{\frac12}\Big(\int |r(u)+r_\delta(u)|^2\rho_0(u)du\Big)^{\frac12}\\
&\le \Big(\int \delta^2L(u)^2\rho_0(u)du\Big)^{\frac12}\Big(\int K_2^2\rho_0(u)du\Big)^{\frac12}\\
&\le \sqrt{K_1}K_2\delta,\quad \delta\in(0,\delta_c).
\end{aligned}$$

And when $\delta\le\tilde\delta_c:=\min\{\frac{Z}{2\sqrt{K_1}K_2},\delta_c\}$, we have

$$Z_\delta\ge Z-|Z-Z_\delta|\ge\frac12 Z.$$

The lemma follows by taking $c_1=\sqrt{K_1}K_2$ and $c_2=\frac12 Z$.

□

Proof of Theorem 3.6. We break the distance into two error parts, one caused by the difference between Z and Z_δ, the other caused by the difference between r and r_δ:

$$\begin{aligned}
d_{\mathtt{Hell}}(\rho,\rho_\delta) &= \frac{1}{\sqrt2}\|\sqrt\rho-\sqrt{\rho_\delta}\|_{L^2}\\
&= \frac{1}{\sqrt2}\Big\|r(u)\sqrt{\frac{\rho_0}{Z}}-r(u)\sqrt{\frac{\rho_0}{Z_\delta}}+r(u)\sqrt{\frac{\rho_0}{Z_\delta}}-r_\delta(u)\sqrt{\frac{\rho_0}{Z_\delta}}\Big\|_{L^2}\\
&\le \frac{1}{\sqrt2}\Big\|r(u)\sqrt{\frac{\rho_0}{Z}}-r(u)\sqrt{\frac{\rho_0}{Z_\delta}}\Big\|_{L^2}+\frac{1}{\sqrt2}\Big\|r(u)\sqrt{\frac{\rho_0}{Z_\delta}}-r_\delta(u)\sqrt{\frac{\rho_0}{Z_\delta}}\Big\|_{L^2}.
\end{aligned}$$

Using Lemma 3.7, for $\delta \in (0, \tilde{\delta}_c)$, we have

$$\begin{aligned}\Big\|r(u)\sqrt{\frac{\rho_0}{Z}} - r(u)\sqrt{\frac{\rho_0}{Z_\delta}}\Big\|_{L^2} &= \Big|\frac{1}{\sqrt{Z}} - \frac{1}{\sqrt{Z_\delta}}\Big|\Big(\int r^2(u)\rho_0(u)du\Big)^{\frac{1}{2}} \\ &= \frac{|Z - Z_\delta|}{(\sqrt{Z} + \sqrt{Z_\delta})\sqrt{Z_\delta}} \\ &\leq \frac{c_1}{2c_2}\delta,\end{aligned}$$

and

$$\Big\|r(u)\sqrt{\frac{\rho_0}{Z_\delta}} - r_\delta(u)\sqrt{\frac{\rho_0}{Z_\delta}}\Big\|_{L^2} = \frac{1}{\sqrt{Z_\delta}}\Big(\int |r(u) - r_\delta(u)|^2\rho_0 du\Big)^{\frac{1}{2}} \leq \sqrt{\frac{K_1}{c_2}}\delta.$$

Therefore

$$d_{\texttt{Hell}}(\rho, \rho_\delta) \leq \frac{1}{\sqrt{2}}\frac{c_1}{2c_2}\delta + \frac{1}{\sqrt{2}}\sqrt{\frac{K_1}{c_2}}\delta = C\delta,$$

with $C = \frac{1}{\sqrt{2}}\frac{c_1}{2c_2} + \frac{1}{\sqrt{2}}\sqrt{\frac{K_1}{c_2}}$ independent of δ. □

3.4 Example

Many practical approximation problems arise from the field of differential equations. Here we consider a simple but typical example where $G(u)$ comes from the solution of an ODE. Let $x(t)$ be the solution to the initial value problem

$$\frac{dx}{dt} = F(x; u), \quad x(0) = 0, \tag{3.1}$$

where $F : \mathbb{R}^J \times \mathbb{R}^N \to \mathbb{R}^J$ is a function such that $F(x; u)$ and the partial Jacobian $D_x F(x; u)$ are uniformly bounded with respect to (x, u), i.e.

$$|F(x; u)|, |D_x F(x; u)| < C, \quad \forall (x, u) \in \mathbb{R}^J \times \mathbb{R}^N,$$

for some constant C, and thus $F(x, u)$ is Lipschitz in x in that

$$|F(x_1; u) - F(x_2; u)| \leq C|x_1 - x_2|, \quad \forall x_1, x_2 \in \mathrm{R}^J.$$

Now consider the inverse problem setting

$$y = G(u) + \eta,$$

where

$$G(u) := x(1) = x(t)|_{t=1},$$

and $\eta \sim N(0, \gamma^2 I_J)$. Assume that in practice the exact mapping $G(u)$ is replaced by some numerical approximation $G_\delta(u)$. In particular, $G_\delta(u)$ is given by using the forward Euler method to solve the ODE (3.1). Define $X_0 = 0$, and

$$X_{k+1} = X_k + \delta F(X_k, u), \quad k \geq 0,$$

where $\delta = \frac{1}{m}$ for some large integer m, then $G_\delta(u)$ is defined as $G_\delta(u) = X_m$.

In what follows, we will prove that $G_\delta(u)$ is uniformly bounded and close to $G(u)$ when δ is small, and then we will use these results to show that Assumption 3.5 is satisfied. Therefore, we can apply Theorem 3.6 to claim that the numerical posterior ρ_δ is a good approximation to the real one ρ in this example.

Define $t_k = k\delta$, $x_k = x(t_k)$. The following lemma gives an estimate on error generated from using the forward Euler method.

Lemma 3.8. *Let $e_k = x_k - X_k$, then we have*

$$|e_k| \le c\delta, \quad 0 \le k \le m,$$

for some constant $c < \infty$; in particular

$$|G(u) - G_\delta(u)| = |e_m| \le c\delta.$$

Proof. For simplicity of exposition we consider the case $N = 1$; the case $N > 1$ is almost identical, simply requiring the integral form for the remainder term in the Taylor expansion. Using Taylor expansion in the case $N = 1$ we have

$$\begin{aligned} x_{k+1} &= x_k + \delta \frac{dx}{dt}(t_k) + \frac{\delta^2}{2}\frac{d^2x}{dt^2}(\xi_k) \\ &= x_k + \delta F(x_k, u) + \frac{\delta^2}{2} D_x F(x(\xi_k), u) F(x(\xi_k), u), \end{aligned}$$

for some $\xi_k \in [t_k, t_{k+1}]$. Then we have

$$\begin{aligned} |e_{k+1}| &= |x_{k+1} - X_{k+1}| \\ &= \left| x_k - X_k + \delta(F(x_k; u) - F(X_k; u)) + \frac{\delta^2}{2} D_x F(x(\xi_k), u) F(x(\xi_k), u) \right| \\ &\le |x_k - X_k| + \delta |F(x_k; u) - F(X_k; u)| + \frac{\delta^2}{2} |D_x F(x(\xi_k), u)||F(x(\xi_k), u)| \\ &\le |e_k| + \delta C |e_k| + \frac{\delta^2}{2} C^2 \end{aligned}$$

Then by the Gronwall inequality, noticing that $|e_0| = 0$, we have

$$\begin{aligned} |e_k| &\le (1 + \delta C)^k |e_0| + \frac{(1 + \delta C)^k - 1}{\delta C} \cdot \frac{\delta^2}{2} C^2 \\ &\le \left((1 + \frac{C}{m})^m - 1 \right) \cdot \frac{C\delta}{2} \\ &\le \frac{(e^C - 1)C}{2} \delta \end{aligned}$$

The lemma follows by taking $c = \frac{(e^C - 1)C}{2}$. □

Lemma 3.9. *$G(u)$ and $G_\delta(u)$ are bounded uniformly with respect to u i.e.*

$$|G(u)|, |G_\delta(u)| < C, \quad \forall u \in \mathrm{R}^N,$$

for some $C \in (0, \infty)$.

Proof. For $G(u)$, we have

$$|G(u)| = |x(1)| = \Big| \int_0^1 F(x(t); u))dt \Big| \leq \int_0^1 |F(x(t); u))|dt \leq C.$$

As for $G_\delta(u)$, we first notice that

$$|X_{k+1}| = |X_k + \delta F(X_k; u)| \leq |X_k| + \delta |F(X_k; u)| \leq |X_k| + \delta C,$$

and by induction we have

$$|X_k| \leq |X_0| + k\delta C = k\delta C,$$

in particular we have

$$|G_\delta(u)| = |X_m| \leq m\delta C = C.$$

□

Next we show that in this example, Assumption 3.5 is satisfied. Recall that $\eta \sim N(0, \gamma^2 I)$, and thus

$$r(u) = \sqrt{\pi(y - G(u))} = \frac{1}{(2\pi)^{\frac{J}{4}} \gamma^{\frac{J}{2}}} \exp\Big(-\frac{1}{4\gamma^2} |y - G(u)|^2 \Big),$$

$$r_\delta(u) = \sqrt{\pi(y - G(u))} = \frac{1}{(2\pi)^{\frac{J}{4}} \gamma^{\frac{J}{2}}} \exp\Big(-\frac{1}{4\gamma^2} |y - G_\delta(u)|^2 \Big).$$

- Assumption 3.5(i): notice that the function e^{-w} is Lipschitz for $w > 0$, with Lipschitz constant 1. Therefore we have

$$\begin{aligned} |r(u) - r_\delta(u)| &\leq \frac{1}{(2\pi)^{\frac{J}{4}} \gamma^{\frac{J}{2}}} \cdot \frac{1}{4\gamma^2} \cdot \big| |y - G(u)|^2 - |y - G_\delta(u)|^2 \big| \\ &= \frac{1}{(2\pi)^{\frac{J}{4}} \gamma^{\frac{J}{2}}} \cdot \frac{1}{4\gamma^2} \cdot |2y - G(u) - G_\delta(u)||G(u) - G_\delta(u)| \\ &\leq \frac{1}{(2\pi)^{\frac{J}{4}} \gamma^{\frac{J}{2}}} \cdot \frac{1}{4\gamma^2} \cdot (2|y| + 2C)c\delta \\ &= \tilde{C}\delta. \end{aligned}$$

  That is to say, Assumption 3.5(i) is satisfied with $L(u) = \tilde{C}$ and $\int_{\mathbb{R}^N} L^2(u)\rho_0(u)du = \tilde{C}^2 < \infty$.

- Assumption 3.5(ii): this assumption is obviously satisfied since

$$r(u) = \frac{1}{(2\pi)^{\frac{J}{4}}\gamma^{\frac{J}{2}}} \exp\Big(-\frac{1}{4\gamma^2}|y - G(u)|^2\Big) \leq \frac{1}{(2\pi)^{\frac{J}{4}}\gamma^{\frac{J}{2}}},$$

$$r_\delta(u) = \frac{1}{(2\pi)^{\frac{J}{4}}\gamma^{\frac{J}{2}}} \exp\Big(-\frac{1}{4\gamma^2}|y - G_\delta(u)|^2\Big) \leq \frac{1}{(2\pi)^{\frac{J}{4}}\gamma^{\frac{J}{2}}}.$$

- Assumption 3.5(iii) The positivity of Z follows from the definition, using the boundedness of G and the fact that ρ_0 is a pdf.

3.5 Discussion and Bibliography

See [39] for more detail on the subject of metrics, and other distance-like functions, on probability measures. See [102, 24] for more detailed discussions on the well-posedness of Bayesian inverse problems, with respect to perturbations in the data; and see [21] for applications concerning numerical approximation of partial differential equations appearing in the forward model. Related results, but using divergences rather than the Hellinger metric, may be found in [80]. The paper [52] contains an interesting set of examples where the Meta Theorem stated at the start of this chapter fails in the sense that, whilst well-posedness holds, the posterior is Hölder with exponent less than one, rather than Lipschitz, with respect to perturbations.

4 Optimization Perspective

In this chapter we explore the properties of Bayesian inversion from the perspective of an optimization problem which corresponds to maximizing the posterior probability, in a sense which we will make precise. We demonstrate the properties of the point estimator resulting from this optimization problem, showing its positive and negative attributes, the latter motivating our work in the following chapter.

4.1 The Setting

Once again we work in the inverse problem setting of finding $u \in \mathbb{R}^N$ from $y \in \mathbb{R}^J$ given by (1.1) with noise $\eta \sim \pi(\cdot)$ and prior $u \sim \rho_0$ as in Assumption 1.1. The posterior $\rho^y(u)$ on $u|y$ is given by Theorem 1.2 and has the form

$$\rho^y(u) = \frac{1}{Z}\pi(y - G(u))\rho_0(u).$$

We may define a loss:

$$\ell(u, y) = -\log \pi(y - G(u)),$$

and a regularizer

$$r(u) = -\log \rho_0(u).$$

When added together these two functions of u comprise an objective function of the form

$$\mathsf{I}(u, y) = \ell(u, y) + r(u).$$

Furthermore

$$\rho^y(u) = \frac{1}{Z}\pi(y - G(u))\rho_0(u) \propto e^{-\mathsf{I}(u,y)}.$$

We see that minimizing the objective function $\mathsf{I}(\cdot, y)$ is equivalent to maximizing the posterior $\rho^y(\cdot)$. We will make this more precise in Theorems 4.6 and 4.7 below.

Definition 4.1. Assume that the infimum of $\mathsf{I}(\cdot, y)$, $\bar{\mathsf{I}}$, is attained at $\bar{u}$. Then $\bar{u}$ is called a *maximum a posteriori (MAP) estimator.*

Example 4.2. Consider the Gaussian setting of Assumption 2.1. Then equation (2.6) shows that the MAP estimator is given by m as defined in Theorem 2.2.

Example 4.3. If $\pi(y) = N(0, \Gamma)$, then $\pi(y - G(u)) \propto \exp(-\frac{1}{2}|y - G(u)|^2_\Gamma)$. So the loss in this case is $\ell(u, y) = \frac{1}{2}|y - G(u)|^2_\Gamma$, a Γ-weighted ℓ_2 loss.

Example 4.4. If we have prior $\rho_0(u) = N(0, C_0)$, then ignoring $u-$independent normalization factors, which appear as constant shifts in $\mathsf{I}(\cdot, y)$, we may take the regularizer as $r(u) = \frac{1}{2}|u|^2_{C_0}$. In particular, if $C_0 = \lambda^{-1}I$, then $r(u) = \frac{\lambda}{2}|u|^2$, an ℓ_2 regularizer.

If we combine Example 4.3 and Example 4.4, then we have objective function

$$\mathsf{I}(u, y) = \frac{1}{2}|y - G(u)|^2_\Gamma + \frac{\lambda}{2}|u|^2,$$

a canonical objective function minimizing Γ-weighted ℓ_2 loss with ℓ_2 regularizer. To connect with future discussions, here λ corresponds to prior precision, and may be learned from data, such as in hierarchical methods.

Example 4.5. As an alternative to the ℓ_2 regularizer, consider $u = (u_1, .., u_N)$ with u_j having prior distribution i.i.d. Laplace. Then $\rho_0(u) \propto \exp(-\lambda \sum_{j=1}^N |u_j|) = \exp(-\lambda |u|_1)$. In this case $r(u) = \lambda |u|_1$, an ℓ_1 regularizer. If we combine this prior with the weighted ℓ_2 loss above then we have objective function

$$\mathsf{I}(u, y) = \frac{1}{2}|y - G(u)|_\Gamma^2 + \lambda |u|_1.$$

Even though this objective function promotes sparse solutions, samples from the underlying posterior distribution are not typically sparse.

4.2 Theory

For any optimization problem for an objective function with a finite infimum, it is of interest to determine whether the infinimum is attained at a finite $u^\star$. Here we consider a class of optimization problems corresponding to our Bayesian posterior distribution that have attainable infinimum. That is, assuming G is continuous, the infinimum of a optimization problem with ℓ_2 loss and ℓ_p regularizer is achieved at some finite $u^\star$, and attainable by the maximum-a-posteriori (MAP) estimator of the corresponding Bayesian problem.

Theorem 4.6 (Attainable MAP Estimator)**.** *Assume that:*

1. $G \in C(\mathbb{R}^N, \mathbb{R}^J)$*, i.e. G is a continuous function;*
2. *the objective function $\mathsf{I}(u, y)$ has ℓ_2 loss as defined in Example 4.3 and ℓ_p regularizer $r(u) = \frac{\lambda}{p}|u|_p^p$, $p \in (0, \infty)$.*

Then the infinimum of $I(u, y)$ is attained at a MAP estimator $u^* \in \mathbb{R}^N$.

Proof. We fix y, so we can denote

$$\mathsf{I}(u) = \mathsf{I}(u, y) = \frac{1}{2}|y - G(u)|_\Gamma^2 + \frac{\lambda}{p}|u|_p^p.$$

Then since $\mathsf{I}(0) < \infty$, and $\mathsf{I}(u) \geq 0$, $\forall u \in \mathbb{R}^N$, we have $0 \leq \bar{\mathsf{I}} = \inf_u I(u) < \infty$.
Let $u^{(n)}$ be a minimizing sequence for $\mathsf{I}(u)$, i.e. $\forall \epsilon > 0$, $\exists M = M(\epsilon) \in \mathbb{N}$, s.t. $\bar{\mathsf{I}} \leq \mathsf{I}(u^{(n)}) \leq \bar{\mathsf{I}} + \epsilon$, $\forall n \geq M$. This implies

$$\frac{\lambda}{p}|u^{(n)}|_p^p \leq \bar{\mathsf{I}} + \epsilon, \qquad \forall n \geq M.$$

Since $\{u^{(n)}\}$ is a bounded sequence, $\exists$ subsequence $\{n_j\}$ s.t. $u^{(n_j)} \to u^*$ for some $u^* \in \mathbb{R}^N$. Then, since I is continuous because G is continuous, we have $\mathsf{I}(u^{(n_j)}) \to \mathsf{I}(u^*)$. This implies $\mathsf{I}(u^*) = \bar{\mathsf{I}}$, because $\bar{\mathsf{I}} \leq \mathsf{I}(u^*) \leq \bar{\mathsf{I}} + \epsilon$, $\forall \epsilon > 0$. We note that since $u^{(n)}$ is a minimizing sequence of $\mathsf{I}(u)$, and minimizing $\mathsf{I}(u)$ corresponds to maximizing ρ^y, u^* is MAP estimator. □

Intuitively the MAP estimator maximizes posterior probability. We make this precise in the following theorem which links the objective function $\mathsf{I}(\cdot, y)$ to small ball probabilities. For economy of notation, we write $\mathsf{I}(u) = \mathsf{I}(u; y)$ in what follows.

Theorem 4.7 (Objective Function and Posterior Probability)**.** *Making the same assumptions as in Theorem 4.6, for any fixed y, define*

$$\mathsf{J}(u; \delta) := \int_{v \in B(u,\delta)} \rho^y(v) dv = \mathbb{P}^{\rho^y}(B(u, \delta)),$$

the probability of a ball with radius δ centered at u. Then, $\forall u_1, u_2 \in \mathbb{R}^N$, we have

$$\lim_{\delta \to 0} \frac{\mathsf{J}(u_1, \delta)}{\mathsf{J}(u_2, \delta)} = e^{I(u_2) - I(u_1)}.$$

Proof.

$$\frac{\mathsf{J}(u_1, \delta)}{\mathsf{J}(u_2, \delta)} = \frac{\int_{|v-u_1|<\delta} e^{-\mathsf{I}(v)} dv}{\int_{|v-u_2|<\delta} e^{-\mathsf{I}(v)} dv}$$

If we Taylor expand $\mathsf{I}(v)$ around u with $|u - v| \leq \delta$, then

$$\mathsf{I}(v) = \mathsf{I}(u) + R(u, v; \delta)$$

for some residual term $R(u, v; \delta)$ which tends to zero with $\delta \to 0$, for u, v in a compact set. Thus $\forall u \in \mathbb{R}^N$, $\forall \epsilon > 0$, we can take $\delta_c = \delta_c(u, \epsilon)$ sufficiently small, such that $|R(u, v; \delta)| < \epsilon$, $\forall \delta \in (0, \delta_c)$. So, $\forall \delta \in (0, \delta_c)$,

$$e^{-\mathsf{I}(u)-\epsilon} \leq e^{-\mathsf{I}(v)} \leq e^{-\mathsf{I}(u)+\epsilon}.$$

This in turn implies

$$C_\delta e^{-\mathsf{I}(u)-\epsilon} = e^{-\mathsf{I}(u)-\epsilon} \int_{|v-u|<\delta} dv \leq \int_{|v-u|<\delta} e^{-\mathsf{I}(v)} dv \leq e^{-\mathsf{I}(u)+\epsilon} \int_{|v-u|<\delta} dv = C_\delta e^{-\mathsf{I}(u)+\epsilon},$$

where $C_\delta = \int_{|v-u|<\delta} dv$ is Lebesgue measure of a ball with radius δ, and is independent of u or v. Applying the preceding result to the ratio of the J's, we have

$$e^{-2\epsilon} e^{\mathsf{I}(u_2)-\mathsf{I}(u_1)} \leq \frac{\mathsf{J}(u_1, \delta)}{\mathsf{J}(u_2, \delta)} \leq e^{2\epsilon} e^{\mathsf{I}(u_2)-\mathsf{I}(u_1)}, \qquad \forall \epsilon > 0.$$

Since $\epsilon \to 0$ as $\delta \to 0$ we obtain the desired result. □

Remark 4.8. Intuitively this theorem shows that maximizing the probability of an infinitesimally small ball is the same as minimizing the objective function $I(\cdot, y)$. This is obvious in finite dimensions, but the proof used generalizes beyond measures which possess a Lebesgue density, and may be used in infinite dimensions. □

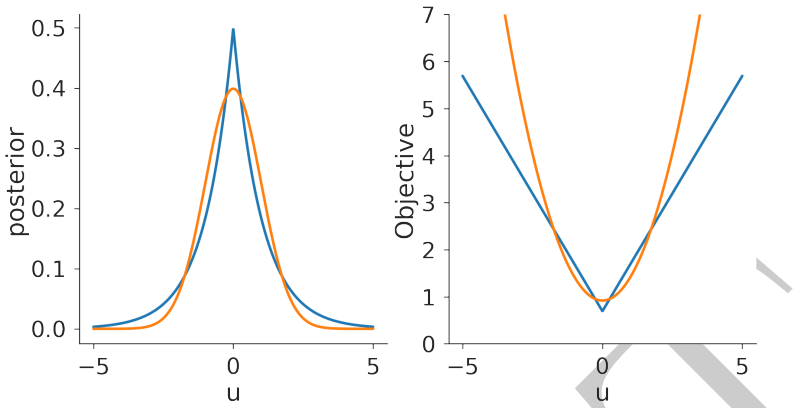


Figure 2 Posterior (left) and objective function (right) for $N(0,1)$ posterior (orange) and Laplace$(0,1)$ posterior (blue).

4.3 Examples

By means of examples we now probe the question of whether or not the MAP estimator captures useful information about the posterior distribution.

Example 4.9. If our posterior is single-peaked, such as a Gaussian or a Laplace distribution, as shown in Figure 2, the MAP estimator, i.e. minimizer of the objective functions, reasonably summarizes where the most likely the true parameter will lie.

Now let's consider one example where a point estimator or a δ-radius ball with $\delta \to 0$ no longer seem to capture what we want:

Example 4.10. If our posterior is rather unevenly distributed, such as a slab-and-spike distribution, as shown in Figure 3, then it is less clear that the MAP estimator usefully summarizes the properties of the posterior. For example, for the case in Figure 3, we may want the solution output of our Bayesian problem to be a weighted average of two Gaussian distributions, or two point estimators each with a separate mean located at one of the two minima of the objective functions, and weight describing the probability mass associated with each of those two points.

Example 4.11. In addition to a multiple-peak posterior, there exist cases where the objective function or the posterior are simply very rough. In these cases, the small-scale roughness should be ignored, while the large-scale variation should be captured. For example in Figure 4, the objective function is very rough, and has a unique minimizer at a point far from 0. However, it also has a larger pattern, that is it tends to be smaller around 0, while larger away from 0. Therefore we see that minimizer of the objective function or the MAP estimator does not capture this large scale pattern. It is arguably the case that $u = 0$ is a better point estimate. An alternative way to view this problem is that there is a natural "temperature" to this problem, in the sense that variations lower than this temperature could be viewed as random noise that do not capture meaningful information for our purpose.

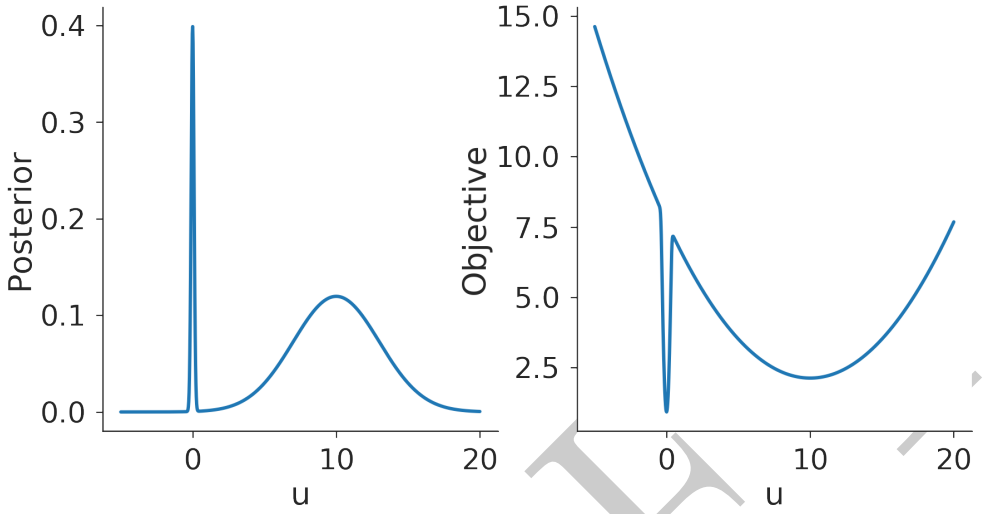


Figure 3 Posterior (left) and objective function (right) for a posterior that is a sum of two Gaussian distributions, $N(0, 0.1^2)$ with probability 0.1 and $N(10, 3^2)$ with probability 0.9.

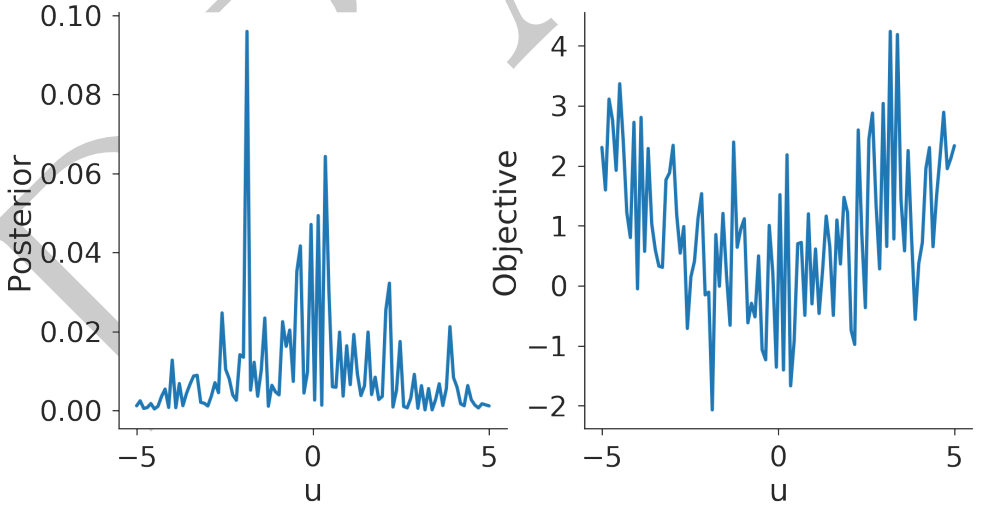


Figure 4 Posterior (left) and objective function (right) from an objective function that is very rough in the small scale, but contains a regular patter on the larger scale. This specific example is generated by white noise summed with a quadratic function for the objective function, and the posterior is computed from the objective function.

The preceding examples suggest that multi-peak distributions, or multi-minimum objective functions, can cause problems for MAP estimation as a point estimator. Here we show that if the dimension N of parameter $u \in \mathbb{R}^N$ is high, then by the sheer fact that the dimension is high, a single point estimator, even if a MAP estimator, is not a good summary of information in the posterior.

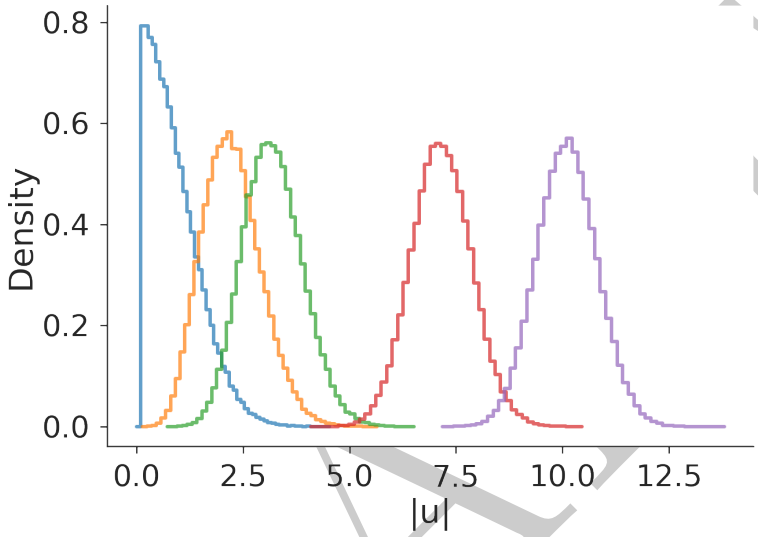


Figure 5 Empirical density of ℓ_2 norm of $N(0, I)$ random vectors for various dimension: $N = 1$ (blue), $N = 5$ (orange), $N = 10$ (green), $N = 50$ (red), and $N = 100$ (purple). The empirical density is obtained from 10000 samples for each distribution.

Example 4.12. We consider what is the "typical size" of a vector u drawn from the standard Gaussian distribution $N(0, I)$, as the dimension increases. In Figure 5 we display the empirical density of the norm of such random vectors. We can see that at low dimensions, such as when $N = 1$, the MAP point, $u = 0$ is highly likely. In higher dimensions, however, the probability for a vector from this distribution to have a small ℓ_2-norm become increasingly small. Indeed, since the sum of n i.i.d. Gaussian random variables with zero mean corresponds to a χ^2 distribution, we can approximate the density accurately for large N. For example, if we look at the probability for the norm to be less than 5, then $\mathbb{P}(|u| < 5)$ is 0.99999943 when $N = 1$, 0.99986 when $N = 5$, 0.99465 when $N = 10$, 0.001192 when $N = 50$, and 1.135×10^{-15} when $N = 100$. So we see that, as the dimension increases, with probability close to 1, a sample from the posterior would have a norm far from 0. Indeed, for $N = 1000$, the 5th and 95th percentiles are respectively 30.3464 and 32.7823. This means when $d = 1000$, we most likely will find a vector with size around 31, not 0. This example therefore suggests for high dimension u, a point estimator simply will not capture what is going on in the density. This effect can also be understood as follows: since the components u_j of u are

i.i.d. standard unit Gaussians we have that, by the strong law of large numbers,

$$\frac{1}{N}\sum_{j=1}^{N} u_j^2 \to 1$$

as $N \to \infty$ almost surely. Thus, with high probability, the ℓ_2-norm is of size $\sqrt{N}$.

The preceding examples demonstrate that MAP estimator should be treated with caution as they may not capture the desired posterior information in many cases. This motivates study of different ways to capture information from the posterior distribution, beyond MAP estimators. One such approach is to try to fit one or several Gaussian distributions to a posterior, by minimizing an appropriate distance-like measure between distributions. This will be discussed in the next chapter.

4.4 Discussion and Bibliography

The optimization perspective on inversion predates the development of the Bayesian approach as a computationa tool, because it is typically far cheaper to implement. The subject of classical regularization techniques for inversion is discussed in [30]. The concept of MAP estimators, which links probability to optimization, is discussed inthe books [62, 103] in the finite dimensional setting. The paper [23] studies this connection precisely: it defines the MAP estimator for infinite dimensional Bayesian inverse problems, and the corresponding variational formulation, in a Gaussian setting. The [48] studies related ideas, but in the non-Gaussian setting. The paper [2] generalizes the variational formulation of MAP estimators to non-Gaussian priors that are sparsity promoting. The paper [110] shows an application of optimization based inversion to a large-scale geophysical application.

5 The Gaussian Approximation

Recall the inverse problem of finding u from y given by (1.1), and the Bayesian formulation which follows from Assumption 1.1. In the previous chapter we explored the idea of obtaining a point estimator for the posterior distribution using an optimization perspective arising from maximizing the posterior pdf. We related this idea to the problem of finding a point estimator that corresponds to finding the center u^* of a ball of radius delta with maximal probability in the limit $\delta \to 0$. Whilst the idea is intuitively appealing, and reduces the complexity of Bayesian inference from determination of an entire distribution to determination of a single point, the approach has a number of limitations, in particular for noisy, multi-peaked or high dimensional posterior distributions; the examples at the end of the previous chapter illustrated this.

In this chapter we again adopt an optimization approach to the problem of Bayesian inference, but instead seek a Gaussian estimator $p \sim N(m, \Sigma)$ of the posterior $\rho^y(u)$. We will find p to minimize the Kullback-Leibler divergence from $\rho^y(u)$; since the Kullback-Leibler divergence is not symmetric this leads to two distinct problems, both of which we will study.

5.1 The Kullback-Leibler Divergence

Definition 5.1. Let $\rho > 0$ and $\rho' > 0$ be two probability distribution on $\mathbb{R}^N$. The *Kullback-Leibler (K-L) divergence*, or *relative entropy*, of ρ with respect to ρ' is

$$\begin{aligned} d_{\mathrm{KL}}(\rho\|\rho') &:= \int_{\mathbb{R}^N} \log\left(\frac{\rho(u)}{\rho'(u)}\right)\rho(u)du \\ &= \mathbb{E}^{\rho}\left[\log\left(\frac{\rho}{\rho'}\right)\right] \\ &= \mathbb{E}^{\rho'}\left[\log\left(\frac{\rho}{\rho'}\right)\frac{\rho}{\rho'}\right] \end{aligned}$$

Remark 5.2. The K-L divergence is not symmetric: in general

$$d_{\mathrm{KL}}(\rho\|\rho') \neq d_{\mathrm{KL}}(\rho'\|\rho).$$

As a consequence, the K-L divergence is not a metric. Nevertheless, it is a convenient quantity to work for at least three reasons: (1) it provides an upper bound for many metrics; (2) its logarithmic structure allows explicit computations that are difficult using actual metrics; (3) it has an information theoretic interpretation, and is part of a *family of divergences* all useful in this context. □

Lemma 5.3. *The K-L divergence is positive and provides an upper bound for both Hellinger and total variation metrics:*

$$d_{Hell}(\rho,\rho')^2 \leq \frac{1}{2}d_{KL}(\rho\|\rho'), \quad d_{TV}(\rho,\rho')^2 \leq d_{KL}(\rho\|\rho').$$

Proof. The second result follows from the first by Lemma 3.2; thus we prove only the first result. Consider the function $g : \mathbb{R}^+ \mapsto \mathbb{R}$ defined by

$$g(x) = x - 1 - \log x.$$

Note that

$$\begin{aligned} g'(x) &= 1 - \frac{1}{x}, \\ g''(x) &= \frac{1}{x^2}, \\ g(\infty) &= g(0) = \infty. \end{aligned}$$

Thus the function is convex on its domain. As the minimum of g is attained at $x = 1$, and as $g(1) = 0$, we deduce that $g(x) \geq 0$ for all $x \in (0, \infty)$. Hence,

$$\begin{aligned} x - 1 &\geq \log x \qquad \forall x \geq 0, \\ \sqrt{x} - 1 &\geq \frac{1}{2} \log x \qquad \forall x \geq 0. \end{aligned}$$

We can use this last inequality to bound the Hellinger distance:

$$\begin{aligned} d_{\texttt{Hell}}(\rho, \rho')^2 &= \frac{1}{2} \int \left(1 - \sqrt{\frac{\rho'}{\rho}}\right)^2 \rho du \\ &= \frac{1}{2} \int \left(1 + \frac{\rho'}{\rho} - 2\sqrt{\frac{\rho'}{\rho}}\right) \rho du \\ &= \int \left(1 - \sqrt{\frac{\rho'}{\rho}}\right) \rho du \leq -\frac{1}{2} \int \log\left(\frac{\rho'}{\rho}\right) \rho du = \frac{1}{2} d_{\texttt{KL}}(\rho \| \rho') \end{aligned}$$

□

5.2 Best Gaussian Fit By Minimizing $d_{\texttt{KL}}(p\|\rho)$

In this section we prove the existence of an approximating Gaussian. We work under the following:

Assumption 5.4. *The posterior distribution is constructed under the following assumptions:*

- *G is bounded and continuous: $G \in C(\mathbb{R}^N, \mathbb{R}^J)$;*
- *we have an ℓ_2-loss function: $\pi(y - G(u)) \propto \exp\left(-\frac{1}{2}|y - G(u)|^2_\Gamma\right)$;*
- *the prior is a centered isotropic Gaussian: $\rho_0 \sim \mathcal{N}(0, \lambda^{-1} I)$.*

Let $\mathcal{A}$ be the set of Gaussian distributions on $\mathbb{R}^N$ given by

$$\mathcal{A} = \{\mathcal{N}(m, \Sigma) : m \in \mathbb{R}^N, \Sigma \in \mathbb{R}^{N \text{x} N} \text{ positive-definite symmetric}\}$$

Theorem 5.5 (Best Gaussian Approximation)**.** *Under Assumption 5.4, there exists at least one probability distribution $p \in \mathcal{A}$ at which the infimum*

$$\inf_{p \in \mathcal{A}} d_{KL}(p\|\rho)$$

is attained.

Proof. The K-L divergence can be computed explicitly as

$$\begin{aligned} d_{\mathtt{KL}}(p\|\rho) &= \mathbb{E}^p \log p - \mathbb{E}^p \log \rho \\ &= \mathbb{E}^p \Big(-\frac{1}{2}|u-m|^2_\Sigma - \frac{1}{2}\log\Big((2\pi)^N \det\Sigma\Big) \; + \; \ell(u;y) + \frac{\lambda}{2}|u|^2 + \log Z \Big) \end{aligned}$$

Note that Z is the normalization constant for ρ and is independent of p and hence of m and Σ. We can represent a given random variable $u \sim p$ by writing $u = m + \Sigma^{1/2}\xi$, where $\xi \sim \mathcal{N}(0, I)$, and hence

$$|u|^2 = |m|^2 + |\Sigma^{1/2}\xi|^2 + 2\langle m, \Sigma^{1/2}\xi\rangle$$

to obtain the functional $\mathsf{J}(m, \Sigma)$ as

$$\begin{aligned} d_{\mathtt{KL}}(p\|\rho) &= \mathbb{E}^p \, \ell(u;y) + \frac{\lambda}{2}|m|^2 + \frac{\lambda}{2}\mathrm{tr}(\Sigma) - \frac{1}{2}\log\det\Sigma - \frac{N}{2} - \frac{N}{2}\log(2\pi) + \log Z \\ &:= \mathsf{J}(m, \Sigma) \end{aligned}$$

As $\mathsf{J}(m, \Sigma) \geq 0$ and $\mathsf{J}(0, I) < \infty$, the infimum of $\mathsf{J}(m, \Sigma)$ over $p \in \mathcal{A}$ is $\bar{\mathsf{J}} \in [0, \infty)$.

Let $(m^{(n)}, \Sigma^{(n)})$ be an infimizing sequence. Then for all $\epsilon > 0$ there exists $M = M(\epsilon)$:

$$0 \leq \bar{\mathsf{J}} \leq \mathsf{J}(m^{(n)}, \Sigma^{(n)}) \leq \bar{\mathsf{J}} + \epsilon \quad \forall n \geq M.$$

The next step in the proof is to show that the sequences $m^{(n)}$ and $\Sigma^{(n)}$ are bounded. To this end let $\sigma_j^{(n)}$ denote the (real and positive) eigenvalues of $\Sigma^{(n)}$, indexed by j. Since the loss function is non-negative we obtain

$$\frac{\lambda}{2}|m^{(n)}|^2 + \sum_{j=1}^{N} \underbrace{\Big(\frac{\lambda}{2}\sigma_j^{(n)} - \frac{1}{2}\log\sigma_j^{(n)}\Big)}_{:= g(\sigma_j^{(n)})} \leq C \quad \forall n \geq M$$

where we have defined the function

$$g(\sigma) = \frac{\lambda\sigma}{2} - \frac{1}{2}\log\sigma.$$

Note further that

$$\begin{aligned} g'(\sigma) &= \frac{\lambda}{2} - \frac{1}{2\sigma} \\ g''(\sigma) &= \frac{1}{2\sigma^2} \\ g(0) &= g(\infty) = \infty \end{aligned}$$

so that g is convex; we denote its lower bound by $-S$. From this it follows that

$$\boxed{\frac{\lambda}{2}|m^{(n)}|^2 \leq C + NS}$$

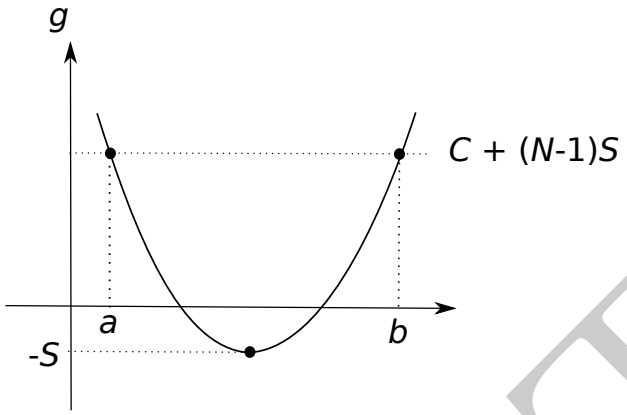


Figure 6 Schematic of g's bounds.

and

$$g(\sigma_k^{(n)}) = \frac{\lambda \sigma_k^{(n)}}{2} - \frac{1}{2}\log \sigma_k^{(n)} \leq C + (N-1)S.$$

Thus

$$\boxed{\sigma_k^{(n)} \in [a,b]} \quad \forall k \in \{1,\ldots,N\};$$

the distinct numbers $a, b \in (0,\infty)$ are displayed in Figure 6, are properties of the function g alone, and in particular are independent of n. These bounds on $m^{(n)}$ and $\sigma_k^{(n)}$ imply that $\exists m^* \in \mathbb{R}^N$, $\sigma_j^* \in [a,b]$, $j = 1,\ldots,N$, such that

$$m^{(n)} \to m^*, \quad \sigma_j^{(n)} \to \sigma_j^*$$

along subsequences $\{n_k\}$ which here we relabel to $\{n\}$. As Σ is an orthonormal similarity transformation of $\Sigma^\star = \mathrm{diag}\{\sigma_1^*,\ldots,\sigma_n^*\}$, then $\Sigma^{(n)} \to \Sigma^*$ along a further subsequence – the orthogonality allows us to extract the further subsequence along which the similarity transformation also converges. Finally, from the dominated convergence theorem, it follows that $\mathsf{J}(m,\Sigma)$ is continuous with respect to (m,Σ), and so

$$\boxed{0 \leq \bar{\mathsf{J}} \leq \mathsf{J}(m^*,\Sigma^*) \leq \bar{\mathsf{J}} + \epsilon} \quad \forall \epsilon > 0.$$

Since ϵ is arbitrary, we deduce that $\mathsf{J}(m^*,\Sigma^*) = \bar{J}$ by sending $\epsilon \to 0$. □

5.3 Best Gaussian Fit By Minimizing $d_{\mathtt{KL}}(\rho\|p)$

Theorem 5.6 (Best Gaussian by Mode Matching). *Assume that $\bar{m} := \mathbb{E}^\rho[u]$ is finite and that $\bar{\Sigma} := \mathbb{E}^\rho[(u-\bar{m})\otimes(u-\bar{m})]$ is positive-definite. Then the infimum*

$$\inf_{p\in\mathcal{A}} d_{KL}(\rho\|p)$$

is attained at the element in $\mathcal{A}$ with mean $\bar{m}$ and covariance $\bar{\Sigma}$.

Proof. We know that

$$d_{\mathtt{KL}}(\rho||p) = -\,\mathbb{E}^{\rho}\left[\log p\right] + \mathbb{E}^{\rho}\left[\log \rho\right] \tag{5.1}$$

Since the second term is constant with respect to p, minimizing $d_{\mathtt{KL}}(\rho\|p)$ is equivalent to minimizing $-\,\mathbb{E}^{\rho}\left[\log p\right]$ which is given by the expression

$$\begin{aligned} -\,\mathbb{E}^{\rho}\left[\log p\right] &= -\,\mathbb{E}^{\rho}\left[\log\left(\frac{1}{\sqrt{(2\pi)^N \det\Sigma}}\exp\left(-\frac{1}{2}|u-m|^2_{\Sigma}\right)\right)\right] \\ &= \frac{1}{2}\,\mathbb{E}^{\rho}\left[|u-m|^2_{\Sigma}\right] + \frac{1}{2}\log\det\Sigma + \frac{N}{2}\log 2\pi \end{aligned}$$

Let $L = \Sigma^{-1}$. Then our task is equivalent to minimizing the following function of m and L:

$$\mathsf{I}(m,L) = \frac{1}{2}\,\mathbb{E}^{\rho}\left[\langle u-m, L(u-m)\rangle\right] - \frac{1}{2}\log\det L$$

First we find the critical points (m, L) for I by taking first order partial derivative with respect to m and L and setting both to zero:

$$\begin{aligned} \partial_m \mathsf{I} &= -\,\mathbb{E}^{\rho}[L(u-m)] = 0; \\ \partial_L \mathsf{I} &= \frac{1}{2}\partial_L(\mathbb{E}^{\rho}[(u-m)\otimes(u-m) : L]) - \frac{1}{2\det L}\partial_L \det L \\ &= \frac{1}{2}\,\mathbb{E}^{\rho}[(u-m)\otimes(u-m)] - \frac{1}{2}L^{-1} = 0; \end{aligned}$$

where we have used the relation $\partial_L \det L = \det L \cdot L^{-T}$ [88]. Solving the above two equations gives us the critical point, expressed in terms of mean and covariance,

$$(\bar{m}, \bar{\Sigma}) = (\mathbb{E}^{\rho}[u], \mathbb{E}^{\rho}[(u-\bar{m})\otimes(u-\bar{m})]).$$

Next we will show $(\bar{m}, \bar{\Sigma}^{-1})$ is a minimizer for I or, equivalently, that $(\bar{m}, \bar{\Sigma})$ is a minimizer for the expression given in (5.1). To this end, if we deal with a vector θ parametrizing the distribution p, it will be sufficient to show that the Hessian of equation (5.1) with respect to θ is positive definite. Recalling the label N of the dimension of the parameter u of unknowns we then have:

$$\begin{aligned} p_{\theta}(u) &= \sqrt{\frac{\det L}{(2\pi)^N}}\exp\left(-\frac{1}{2}(u-m)^T L(u-m)\right) \\ &= \sqrt{\frac{\det L}{(2\pi)^N}}\exp\left[-\left(\frac{1}{2}u^T L u - m^T L u + \frac{1}{2}m^T L m\right)\right] \\ &= \sqrt{\frac{\det L}{(2\pi)^N}}\exp\left(-\frac{1}{2}m^T L m\right)\exp\left(-\frac{1}{2}u^T L u + m^T L u\right) \end{aligned}$$

Utilizing vectorization of matrices we may write:

$$u^T L u = L : uu^T = [\mathrm{vec}(L)]^T \mathrm{vec}(uu^T)$$
$$m^T L u = (Lm)^T u$$
$$\Rightarrow -\frac{1}{2} u^T L u + m^T L u = \begin{bmatrix} Lm \\ -\frac{1}{2}\mathrm{vec}(L) \end{bmatrix}^T \begin{bmatrix} u \\ \mathrm{vec}(uu^T) \end{bmatrix}$$

Then we let:

$$\theta = \begin{bmatrix} Lm \\ -\frac{1}{2}\mathrm{vec}(L) \end{bmatrix}, \qquad T(u) = \begin{bmatrix} u \\ \mathrm{vec}(uu^T) \end{bmatrix}$$

The pdf $p_\theta(u)$ can then be written in the following form:

$$p_\theta(u) = \frac{1}{Z(\theta)} \exp\Big(\theta^T T(u)\Big) \tag{5.2}$$
$$\text{with} \quad Z(\theta) = \int_{\mathbb{R}^N} \exp\Big(\theta^T T(u)\Big) du \tag{5.3}$$

Using equation (5.2) we can rewrite equation (5.1) as:

$$H(\theta) = d_{\mathtt{KL}}(\rho || p_\theta) = -\theta^T \mathbb{E}^\rho[T(u)] + \log(Z(\theta)) + \mathbb{E}^\rho[\rho(u)] \tag{5.4}$$

Notice that

$$\nabla_\theta \log(Z(\theta)) = \frac{1}{Z(\theta)} \int_{\mathbb{R}^N} \nabla_\theta \Big[\exp\Big(\theta^T T(u)\Big)\Big] du \tag{5.5}$$
$$= \frac{1}{Z(\theta)} \int_{\mathbb{R}^N} T(u) \exp\Big(\theta^T T(u)\Big) du \tag{5.6}$$
$$= \mathbb{E}^{p_\theta}[T(u)]] \tag{5.7}$$

Therefore we can calculate the gradient and Hessian of $H(\theta)$ as follows:

$$\nabla_\theta H(\theta) = -\mathbb{E}^\rho[T(u)] + \mathbb{E}^{p_\theta}[T(u)], \tag{5.8}$$
$$\Big[\nabla^2_\theta H(\theta)\Big]_{ij} = \frac{\partial \log(Z(\theta))}{\partial \theta_i \theta_j} = \frac{\partial}{\partial \theta_j}\Big(\frac{1}{Z(\theta)} \int_{\mathbb{R}^N} T_i(u) e^{\theta^T T(u)} du\Big) \tag{5.9}$$
$$= \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta_j} \int_{\mathbb{R}^N} T_i(u) e^{\theta^T T(u)} dx + \int_{\mathbb{R}^N} T_i(u) e^{\theta^T T(u)} du \cdot \frac{\partial}{\partial \theta_j} Z(\theta)^{-1} \tag{5.10}$$
$$= \frac{1}{Z(\theta)} \int_{\mathbb{R}^N} T_i(u) T_j(u) e^{\theta^T T(u)} du \tag{5.11}$$
$$- \frac{1}{Z(\theta)^2} \int_{\mathbb{R}^N} T_i(u) e^{\theta^T T(u)} du \cdot \int_{\mathbb{R}^N} T_j(u) e^{\theta^T T(u)} du \tag{5.12}$$
$$= \mathbb{E}^{p_\theta}[T_i T_j] - \mathbb{E}^{p_\theta}[T_i]\, \mathbb{E}^{p_\theta}[T_j] \tag{5.13}$$
$$= [\mathrm{Cov}^{p_\theta}(T)]_{ij}. \tag{5.14}$$

Therefore the Hessian of the objective function is the covariance matrix of $T(u)$ under the multivariate normal distribution p_θ and by construction, is positive semi-definite. Therefore $d_{\mathtt{KL}}(\rho || p_\theta)$ is convex in θ and the critical point $(\bar{m}, \bar{\Sigma})$ is a corresponding minimizer. □

Remark 5.7. Notice that the preceding proof of convexity holds for any distribution p that can alternatively be parametrized by a vector θ with the general form of equation (5.2). In particular for such problems the convexity of $d_{\mathtt{KL}}(\rho\|p_\theta)$ follows, ensuring the existence of a minimizer. In fact, equation (5.2) is a reduced form of the following more general expression:

$$p_\theta(u) = h(u)\exp\Big(\theta^T T(u) - A(\theta)\Big) \tag{5.15}$$

$$\text{with} \qquad A(\theta) = \log\left[\int_{\mathbb{R}^N} h(u)\exp\Big(\theta^T T(u)\Big)du\right] \tag{5.16}$$

Since $h(u)$ is independent of θ, the conclusion of the previous theorem carry over to distributions with the form of (5.15). Such distributions are said to termed the *exponential family* in the statistics literature, and θ is called the natural parameter, $T(u)$ the sufficient statistic, $h(u)$ the base measure, and $A(\theta)$ the log-partition. The Gaussian distribution is a special case in which $h(u)$ is constant with respect to u. In what follows we show some examples of other exponential family members. □

Example 5.8. The Bernoulli distribution

Let u be a random variable following the Bernoulli distribution with $\mathbb{P}(u=1) = q$ and $\mathbb{P}(u=0) = 1-q$. Then the point mass function describing the distribution of u is:

$$\begin{aligned} p(u;q) &= q^u(1-q)^{1-u} \\ &= \exp\left[u\log\left(\frac{q}{1-q}\right) + \log(1-q)\right] \\ &= (1-q)\exp\left[u\log\left(\frac{q}{1-q}\right)\right] \end{aligned}$$

Therefore $\theta = \log\Big(\frac{q}{1-q}\Big)$, $T(u) = u$, $A(\theta) = \log(1+e^\theta)$, and $h(u) = 1$.

Example 5.9. The Poisson distribution

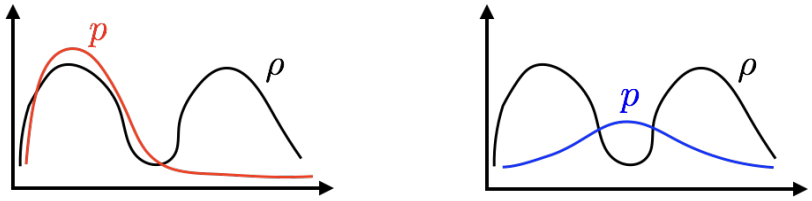
The the point mass function describing the distribution of u is, in this case,:

$$\begin{aligned} p(u;\lambda) &= \frac{\lambda^u e^{-\lambda}}{u!} \\ &= \frac{1}{u!}e^{u\log\lambda - \lambda} \end{aligned}$$

Therefore $\theta = \log\lambda$, $T(u) = u$, $A(\theta) = e^\theta$ and $h(u) = \frac{1}{u!}$.
For more information about the properties and examples of exponential family distributions, we refer to [86].

5.4 Comparison between $d_{\mathtt{KL}}(\rho||p)$ and $d_{\mathtt{KL}}(p||\rho)$

It is instructive to compare the two different minimization problems, both leading to a "best Gaussian", that we described in the preceding two sections. We write the two



(a) Minimizing $d_{\mathtt{KL}}(p||\rho)$ **(b)** Minimizing $d_{\mathtt{KL}}(\rho||p)$

Figure 7 (a) Minimizing $d_{\mathtt{KL}}(p||\rho)$ can lead to serious information loss while (b) minimizing $d_{\mathtt{KL}}(\rho||p)$ ensures a comprehensive consideration of all components of ρ.

relevant divergences as follows and then explain the nomenclature:

$$d_{\mathtt{KL}}(p||\rho) = \mathbb{E}^p\left[\log\left(\frac{p}{\rho}\right)\right] = \mathbb{E}^p\left[\log p\right] - \mathbb{E}^p\left[\log \rho\right] \quad \text{"Mode-seeking"}$$
$$d_{\mathtt{KL}}(\rho||p) = \mathbb{E}^\rho\left[\log\left(\frac{\rho}{p}\right)\right] = \mathbb{E}^\rho\left[\log \rho\right] - \mathbb{E}^\rho\left[\log p\right] \quad \text{"Mean-seeking"}$$

Note that when minimizing $d_{\mathtt{KL}}(p\|\rho)$ we want $\log\frac{p}{\rho}$ to be small, which can happen when $p \simeq \rho$ or $p \ll \rho$. This illustrates the fact that minimizing $d_{\mathtt{KL}}(p||\rho)$ may miss out components of ρ. For example, in Figure 7(a) ρ is a bi-modal like distribution but minimizing $d_{\mathtt{KL}}(p||\rho)$ over Gaussians p can only give a single mode approximation which is achieved by matching one of the modes; we may think of this as "mode-seeking". In contrast, when minimizing $d_{\mathtt{KL}}(\rho||p)$ over Gaussians p we want $\log\frac{\rho}{p}$ to be small where p appears as the denominator. This implies that wherever ρ has some mass we must let p also have some mass there in order to keep $\frac{\rho}{p}$ as close as possible to one. Therefore the minimization is carried out by allocating the mass of p in a way such that on average the divergence between p and ρ attains its minimum as shown in Figure 7(b); hence the label "mean-seeking." Different applications will favor different choices between the mean and mode seeking approaches to Gaussian approximation.

5.5 Variational Formulation of Bayes Theorem

This chapter has been concerned with finding the best Gaussian approximation to a measure with respect to KL divergences. Bayes Theorem 1.2 itself can be formulated through a closely related minimization principle. Consider posterior $\rho(u)$ in the following form:

$$\rho(u) = \frac{1}{Z}\exp(-l(u;y))\rho_0(u)$$

where $\rho_0(u)$ is the prior, $l(u;y)$ the negative log likelihood and Z the normalization constant. We also assume the following:

Assumption 5.10. *The likelihood and prior satisfy:*

- $l : \mathbb{R}^N \times \mathbb{R}^J \to [0,\infty]$;
- $\rho_0 : \mathbb{R}^N \to \mathbb{R}^+$ *is everywhere strictly positive on* $\mathbb{R}^N$.

Then we can express $d_{\mathtt{KL}}(p||\rho)$ in terms of the prior:

$$\begin{aligned} d_{\mathtt{KL}}(p||\rho) &= \int_{\mathbb{R}^N} \log\left(\frac{p}{\rho}\right) p du \\ &= \int_{\mathbb{R}^N} \log\left(\frac{p}{\rho_0}\frac{\rho_0}{\rho}\right) p du \\ &= \int_{\mathbb{R}^N} \log\left(\frac{p}{\rho_0}\exp(l(u;y))\, Z\right) p du \\ &= d_{\mathtt{KL}}(p||\rho_0) + \mathbb{E}^p[l(u;y)] + \log Z \end{aligned}$$

If we define

$$J(p) = d_{\mathtt{KL}}(p||\rho_0) + \mathbb{E}^p[l(u;y)]$$

then we have the following:

Theorem 5.11 (Bayes Theorem as an Optimization Principle)**.** *The posterior measure ρ is given by the following minimization principle:*

$$\rho = \mathrm{argmin}_{p\in\mathcal{P}} J(p),$$

where $\mathcal{P}$ contains all pd on $\mathbb{R}^N$.

Proof. Since Z is the normalization constant for ρ and is independent of p, the minimizer of $d_{\mathtt{KL}}(p\|\rho)$ will also be the minimizer for $J(p)$ and so it suffices to show that the global minimizer of $d_{\mathtt{KL}}(p\|\rho)$ is attained at $p = \rho$. Let $h = x\log x$, noting that this is convex on $x > 0$. Then we have, by Jensen's inequality,

$$\begin{aligned} d_{\mathtt{KL}}(p\|\rho) &= \int_{\mathbb{R}^N} h\left(\frac{p}{\rho}\right)\rho du \\ &\geq h\left(\int_{\mathbb{R}^N} \frac{p}{\rho}\rho du\right) = h(1) = 0. \end{aligned}$$

Note that $p = \rho \Rightarrow d_{\mathtt{KL}}(p\|\rho) = 0$. Thus the choice $p = \rho$ achieves the global minimum. Since h is not affine in x equality in the Jensen lower bound holds if and only if $\frac{p}{\rho}$ is a constant almost everywhere. Given ρ and p are both pdfs on $\mathbb{R}^N$ this further implies that $p = \rho$. Therefore we must have that the minimum of $d_{\mathtt{KL}}(p\|\rho)$ is attained if and only if $\rho = p$. Hence the desired conclusion. □

5.6 Discussion and Bibliography

For definition of the K-L divergence, and upper-bounds in terms of probability metrics, see [39]. For a basic introduction to variational methods, and the moment-matching version of Gaussian approximation, see [13]. For the problems of finding a Gaussian approximation of a general finite dimensional probability distribution see [76]. For infinite dimensional formulations of the problem the reader is referred to [89] and the companion paper [90]. The approximation in Theorem 5.5 consists of a single Gaussian distribution. If the posterior has more than one mode, a single Gaussian may not

be appropriate. For an approximation composed of Gaussian mixtures, the reader is referred to [76]. The formulation of Bayes Theorem as an optimization principle is well-known and widely used in the machine learning community where it goes under the name "variational Bayes"; see the book [77] and the paper [12] for clear expositions of this subject.

6 Importance Sampling

In this chapter, we introduce Monte Carlo sampling and importance sampling. They are two general techniques for estimating expectations computed with respect to a particular distribution, by generating samples from (a possibly different) distribution; they may also be viewed as approximations of a given measure via sums of Dirac measures.

Throughout this chapter we focus interest on computing expectations with respect to a probability distribution with pdf ρ given in the form

$$\rho(u) = \frac{1}{Z} g(u)\rho_0(u), \quad Z = \rho_0(g). \tag{6.1}$$

Monte Carlo sampling will use samples from ρ itself; importance sampling will use samples from ρ_0. A particular application of this setting is where ρ_0 is the prior, ρ the posterior and g is given by Bayes Theorem 1.2. Let $\varphi : \mathbb{R}^N \longrightarrow \mathbb{R}$ denote the functional whose expectation we are interested in computing. For any pdf ν on $\mathbb{R}^N$ we write

$$\nu(\varphi) = \mathbb{E}^{\nu}[\varphi(u)] = \int_{\mathbb{R}^N} \varphi(u)\nu(u)du. \tag{6.2}$$

In this chapter we will generalize the concept of pdf to include Dirac mass distributions. A Dirac mass at v will be viewed as having pdf $\delta(\cdot - v)$ where $\delta(\cdot)$ integrates to one and takes the value zero everywhere except at the origin.

6.1 Monte Carlo Sampling

If we have M random samples $u^{(1)}, ..., u^{(M)}$, generated i.i.d according to ρ, then we can estimate ρ by the Monte Carlo estimator ρ^M_{MC}, which is defined as

$$\rho^M_{MC} := \frac{1}{M} \sum_{m=1}^{M} \delta(u - u^{(m)}). \tag{6.3}$$

And this gives rise to the following estimator of $\rho(\varphi)$:

$$\rho^M_{MC}(\varphi) := \frac{1}{M} \sum_{m=1}^{M} \varphi(u^{(m)}), \ u^{(m)} \sim \rho \text{ i.i.d.}$$

Theorem 6.1 (Monte Carlo Error). *Let $\varphi : \mathbb{R}^N \longrightarrow \mathbb{R}$. We have*

$$\sup_{|\varphi|\leq 1} \left| \mathbb{E}\left[\rho^M_{MC}(\varphi) - \rho(\varphi)\right] \right| = 0,$$
$$\sup_{|\varphi|\leq 1} \left| \mathbb{E}\left[\left(\rho^M_{MC}(\varphi) - \rho(\varphi)\right)^2\right] \right| \leq \frac{1}{M}.$$

Proof. Define

$$\bar{\varphi}(u) = \varphi(u) - \rho(\varphi).$$

To prove the first result, namely that the estimator is unbiased, note that

$$\begin{aligned}
&\mathbb{E}\left[\rho_{MC}^{M}(\varphi) - \rho(\varphi)\right] \\
&= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\varphi\left(u^{(m)}\right) - \rho(\varphi)\right] \\
&= \frac{1}{M}\sum_{m=1}^{M} \left(\rho(\varphi) - \rho(\varphi)\right) \\
&= \frac{1}{M}\cdot 0 = 0.
\end{aligned}$$

Therefore the supremum of its absolute value is also zero. For the second result, which estimates the variance of the error in the estimator, we observe that $\mathbb{E}\left[\bar{\varphi}\right] = 0$ and, then,

$$\begin{aligned}
\mathbb{E}\left[\left(\rho_{MC}^{M}(\varphi) - \rho(\varphi)\right)^2\right] &= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{n=1}^{M} \mathbb{E}\left[\bar{\varphi}\left(u^{(m)}\right)\bar{\varphi}\left(u^{(n)}\right)\right] \\
&= \frac{1}{M^2}\sum_{m=1}^{M} \mathbb{E}\left[\bar{\varphi}\left(u^{(m)}\right)^2\right] \\
&= \frac{1}{M}\,\mathbb{E}\left[\bar{\varphi}\left(u^{(1)}\right)^2\right] = \frac{1}{M}\mathrm{Var}_{\rho}[\varphi]
\end{aligned}$$

since $u^{(m)}$ are i.i.d. In particular we have

$$\mathbb{E}\left[\left(\rho_{MC}^{M}(\varphi) - \rho(\varphi)\right)^2\right] = \frac{1}{M}\mathrm{Var}_{\rho}[\varphi] \leq \frac{1}{M}\rho(\varphi^2) \tag{6.4}$$

since

$$\mathrm{Var}_{\rho}[\varphi] = \rho(\varphi^2) - \rho(\varphi)^2 \leq \rho(\varphi^2).$$

Therefore

$$\sup_{|\varphi|\leq 1}\left|\mathbb{E}\left[\left(\rho_{MC}^{M}(\varphi) - \rho(\varphi)\right)^2\right]\right| = \sup_{|\varphi|\leq 1}\left|\frac{1}{M}\mathrm{Var}_{\rho}[\varphi]\right| \leq \frac{1}{M}.$$

□

The theorem shows that the Monte Carlo estimator ρ_{MC}^{M} is an unbiased approximation for the posterior ρ and that, by choosing M large enough, expectation of any bounded function φ can in principle be approximated by Monte Carlo sampling to arbitrary accuracy. Furthermore, although the convergence is slow with respect to M there is no dependence on the dimension of the problem or on the properties of φ, other than its supremum.

Example 6.2 (Approximation of an Integral)**.** Let $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ be a sigmoid function defined on $\mathbb{R}$ and shown in Figure 8(a) below as the blue solid curve. Let $u \sim \rho = 0.1\,\mathcal{N}(-5,1) + 0.9\,\mathcal{N}(5,1)$ be a Gaussian mixture consisting of two Gaussian distributions. We wish to approximate the expected value of $\varphi(u) \times \mathbb{I}_{[a,b]}(u)$ and $\mathbb{I}_{[a,b]}(u) = \begin{cases} 1 & \text{if } u \in [a,b] \\ 0 & \text{otherwise} \end{cases}$. We

use Monte Carlo sampling to generate M random samples $u^{(1)}, ..., u^{(M)}$ and compute the error between the actual integral and Monte Carlo estimator. The integral and estimator are in the form:

$$\rho(\varphi) = \int_a^b \varphi(u)\rho(u)\,\mathrm{d}u,$$
$$\rho_{MC}^{M}(\varphi) = \frac{1}{M}\sum_{m=1}^{M} \varphi(u^{(m)})\mathbb{I}_{[a,b]}(u^{(m)}).$$

The results of a set of numerical experiments, for different M, are shown in Figure 8(b), in the case where $a = -5, b = 5$. A subset of the 100 samples used is displayed in Figure 8(a).

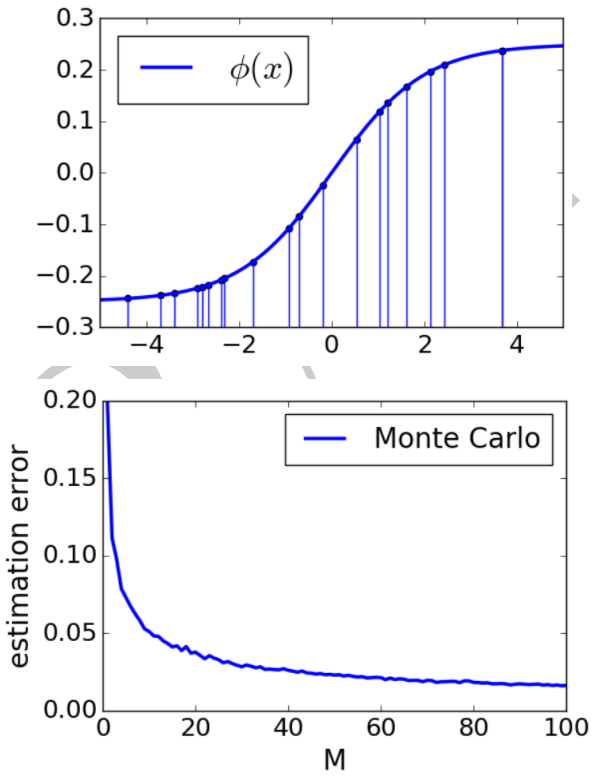


Figure 8 Large sampling number reduces the estimation error by Monte Carlo method.

6.2 Importance Sampling

Standard Monte Carlo sampling is good if we can sample from the desired *target* distribution ρ. When it is hard to sample from ρ, we can draw samples from another *proposal* distribution ρ_0 instead. We then need to evaluate g in (6.1) at each sample and use it as an importance weight in the approximation of the desired distribution. This is the idea of importance sampling.

If we have M random samples $u^{(1)}, ..., u^{(M)}$, generated i.i.d according to ρ_0, then we can estimate ρ by ρ_{IS}^M, defined as

$$\rho_{IS}^M := \sum_{m=1}^{M} w_m \delta(u - u^{(m)}),$$

where

$$w_m = \frac{g\Big(u^{(m)}\Big)}{\sum_{l=1}^{M} g\big(u^{(l)}\big)}.$$

The resulting approximation of $\rho(\varphi)$ is given by

$$\rho_{IS}^M(\varphi) := \sum_{m=1}^{M} w_m \varphi\Big(u^{(m)}\Big),\ u^{(m)} \sim \rho \text{ i.i.d},$$

Theorem 6.3 (Importance Sampling Error). *Let $\varphi : \mathbb{R}^N \longrightarrow \mathbb{R}$. We have*

$$\sup_{|\varphi|\leq 1} \Big|\mathbb{E}\Big[\rho_{IS}^M(\varphi) - \rho(\varphi)\Big]\Big| \leq \frac{2r}{M},$$
$$\sup_{|\varphi|\leq 1} \Big|\mathbb{E}\Big[\Big(\rho_{IS}^M(\varphi) - \rho(\varphi)\Big)^2\Big]\Big| \leq \frac{4r}{M},$$

where

$$r = \frac{\rho_0(g^2)}{\rho_0(g)^2}.$$

Proof. Given

$$\rho(u) = \frac{1}{Z} g(u)\rho_0(u) = \frac{1}{\rho_0(g)} g(u)\rho_0(u),$$

we can rewrite

$$\rho(\varphi) = \frac{\rho_0(g\varphi)}{\rho_0(g)} \simeq \frac{\rho_{0,MC}^M(g\varphi)}{\rho_{0,MC}^M(g)} = \rho_{IS}^M(\varphi).$$

Then we have

$$\begin{aligned}\rho_{IS}^M(\varphi) - \rho(\varphi) &= \frac{\rho_{0,MC}^M(g\varphi)}{\rho_{0,MC}^M(g)} - \frac{\rho_0(g\varphi)}{\rho_0(g)} \\ &= \frac{\rho_{IS}^M(\varphi)\Big(\rho_0(g) - \rho_{0,MC}^M(g)\Big)}{\rho_0(g)} - \frac{\Big(\rho_0(g\varphi) - \rho_{0,MC}^M(g\varphi)\Big)}{\rho_0(g)}.\end{aligned}$$

The expectation of the second term is zero and hence

$$\begin{aligned}\Big|\mathbb{E}\Big[\rho_{IS}^M(\varphi) - \rho(\varphi)\Big]\Big| &= \frac{1}{\rho_0(g)}\Big|\mathbb{E}\Big[\rho_{IS}^M(\varphi)\Big(\rho_0(g) - \rho_{0,MC}^M(g)\Big)\Big]\Big| \\ &\leq \frac{1}{\rho_0(g)}\Big|\mathbb{E}\Big[\Big(\rho_{IS}^M(\varphi) - \rho(\varphi)\Big)\Big(\rho_0(g) - \rho_{0,MC}^M(g)\Big)\Big]\Big|,\end{aligned}$$

since $\mathbb{E}\left[\rho_0(g) - \rho^M_{0,MC}(g)\right] = 0$. Using the Cauchy-Schwarz inequality, the second result from this theorem (whose proof follows) and (6.4) from the proof of Theorem 6.1 we have $\forall|\varphi| \leq 1$,

$$\begin{aligned}\left|\mathbb{E}\left[\rho^M_{IS}(\varphi) - \rho(\varphi)\right]\right| &\leq \frac{1}{\rho_0(g)}\left(\mathbb{E}\left[\left(\rho^M_{IS}(\varphi) - \rho_0(\varphi)\right)^2\right]\right)^{1/2}\left(\mathbb{E}\left[\left(\rho_0(g) - \rho^M_{0,MC}(g)\right)^2\right]\right)^{1/2} \\ &\leq \frac{1}{\rho_0(g)}\left(\frac{4r}{M}\right)^{1/2}\left(\frac{\rho_0(g^2)}{M}\right)^{1/2} = \frac{2r}{M}.\end{aligned}$$

We now prove the second result. We use the splitting of $\rho^M_{IS}(\varphi) - \rho(\varphi)$ into the sum of two terms as derived above. Using (6.4), together with the facts that

$$\left|\mathbb{E}\left[(X-Y)^2\right]\right| \leq 2\Big(\mathbb{E}\left[X^2\right] + \mathbb{E}\left[Y^2\right]\Big),$$

and $\forall|\varphi| \leq 1, |\rho^M_{IS}(\varphi)| \leq 1$, we have $\forall|\varphi| \leq 1$,

$$\begin{aligned}&\left|\mathbb{E}\left[\left(\rho^M_{IS}(\varphi) - \rho(\varphi)\right)^2\right]\right| \\ &\leq \frac{2}{\rho_0(g)^2}\left(\mathbb{E}\left[\left(\rho^M_{IS}(\varphi)\right)^2\left(\rho_0(g) - \rho^M_{0,MC}(g)\right)^2\right] + \mathbb{E}\left[\left(\rho_0(g\varphi) - \rho^M_{0,MC}(g\varphi)\right)^2\right]\right) \\ &\leq \frac{2}{\rho_0(g)^2}\left(\mathbb{E}\left[\left(\rho_0(g) - \rho^M_{0,MC}(g)\right)^2\right] + \mathbb{E}\left[\left(\rho_0(g\varphi) - \rho^M_{0,MC}(g\varphi)\right)^2\right]\right) \\ &= \frac{2}{\rho_0(g)^2 M}\left(\mathrm{Var}_{\rho_0}[g] + \mathrm{Var}_{\rho_0}[g\varphi]\right) \\ &\leq \frac{2}{\rho_0(g)^2 M}\Big(\rho_0(g^2) + \rho_0(g^2\varphi^2)\Big) \\ &\leq \frac{4\rho_0(g^2)}{\rho_0(g)^2 M} = \frac{4r}{M}.\end{aligned}$$

Therefore,

$$\sup_{|\varphi|\leq 1}\left|\mathbb{E}\left[\left(\rho^M_{IS}(\varphi) - \rho(\varphi)\right)^2\right]\right| \leq \frac{4r}{M}.$$

□

This theorem shows that, unlike Monte Carlo, the Importance Sampling estimator ρ^M_{IS} is a biased approximation for the posterior ρ. The rate of convergence of the bias is twice that of the standard deviation, however. As for Monte Carlo the rate of convergence of the variance of the error is governed by the inverse of M, but is independent of the dimension and of the specific φ, provided it is bounded by one. However for Importance Sampling to be accurate (with a limited number of samples M) it is important that r is not too large. The size of r measures, intuitively, how close ρ_0 is to ρ. Indeed we have the following result which, in words, says that the minimum value of r, namely unity, is achieved if and only if ρ and ρ_0 are the same probability density.

Lemma 6.4.

$$r = \frac{\rho_0(g^2)}{\rho_0(g)^2} \geq 1, and\ r = 1 \iff g \equiv const.$$

Proof. By definition of variance, we have

$$\mathrm{Var}_{\rho_0}[g] = \rho_0\Big(g^2\Big) - \rho_0(g)^2 \geq 0.$$

Thus $\rho_0(g^2) \geq \rho_0(g)^2$, and hence $r \geq 1$. Furthermore

$$r = 1 \iff \mathrm{Var}_{\rho_0}[g] = 0 \iff g \equiv const.$$

□

Example 6.5 (Change of Measurement)**.** We consider a similar set-up as in Example 6.2, integrating a sigmoid function, shown in blue in Figure 9, with respect to a probability measure ρ which is bimodal, shown in red in Figure 9; we again restrict the support of the desired integral. We estimate the integral using Importance Sampling based on M random samples $u^{(1)}, ..., u^{(M)}$ from measure $\rho_0 = \mathcal{N}(\mu, \sigma^2)$, shown in green in Figure 9. The estimator of the integral is given by

$$\rho_{IS}^M(\varphi) = \sum_{m=1}^{M} w_m \varphi(u^{(m)}) \mathbb{I}_{[a,b]}(u^{(m)})$$
$$w_m = \frac{g(u^{(m)})}{\sum_{l=1}^{M} g(u^{(m)})}.$$

Here g is a function proportional to the ratio of the densities of ρ and ρ_0. If $\rho(u^{(m)}) > \rho_0(u^{(m)})$, the samples should have been denser, so we raise the weight on $\varphi(u^{(m)})$ in proportion to $\frac{\rho(u^{(m)})}{\rho_0(u^{(m)})} > 1$. If $\rho(u^{(m)}) < \rho_0(u^{(m)})$, the samples should have been less dense, so we lower the weight on $\varphi(u^{(m)})$ in proportion to $\frac{\rho(u^{(m)})}{\rho_0(u^{(m)})} < 1$.

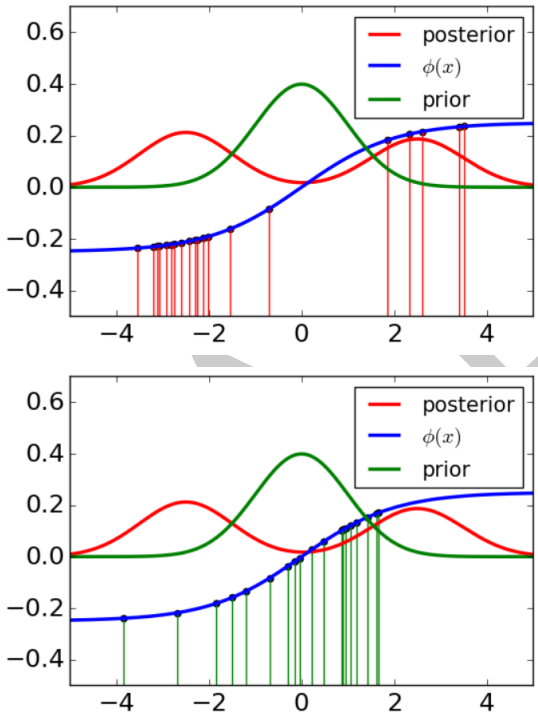


Figure 9 IS is a change of measurement via the importance weights. The red curve shows bimodal distribution of ρ and the green curve shows the Gaussian distribution of ρ_0. The blue curve is the function to be integrated, on its support $[-5, 5]$. The upper figure shows samples from the posterior ρ itself; these would be used for Monte Carlo sampling; the lower curve shows samples from the prior ρ_0, as used for Importance Sampling. The importance weights capture and compensate for the difference of sampling from these two distributions.

6.3 Discussion and Bibliography

[46] is a classic reference on Monte Carlo method and includes discussion of several variance-reduction techniques. The chapter notes [6] give a comparison of Monte Carlo and Importance Sampling with examples. The paper [61] further explores advanced Importance Sampling via adaptive algorithms. When M is large enough, ρ_{MC}^{M} arising from a Monte Carlo simulation should be close to ρ [32]. In practice, all probabilities, integrals and summations can be approximated by the Monte Carlo method [104]. A review of importance sampling, from the perspective of filtering and sequential importance resampling, may be found in [4]; the proofs in this chapter closely follow the presentation in that paper. The subject of multi level Monte Carlo (MLMC) has made the use of Monte Carlo methods practical in new areas of application; see [66] for an overview. The methodology applies when approximating expectations over infinite dimensional spaces, and distributes the computational budget over different levels of approximation, with the goal of optimizing the cost per unit error, noting that the latter balances sampling and approximation based sources.

7 Monte Carlo Markov Chain

In this chapter we focus on Monte Carlo Markov Chain (MCMC) as a methodology to sample from a given *target distribution* ρ on $\mathbb{R}^n$. As with Monte Carlo and importance sampling, the method may be viewed as approximating the posterior distribution by a sum of Dirac measures. After some discussion of the general methodology, we will specify to the case where ρ is a posterior distribution given via Bayes theorem from the product of the likelihood and the prior distributions. We will study a specific MCMC method which uses these two ingredients, the prior and the likelihood, as part of its design.

7.1 The Idea Behind MCMC

The idea of MCMC is to construct a Markov Chain that has the target distribution as its equilibrium distribution. [1] The MCMC algorithm will draw a series of samples $\{u_k\}_{k=1}^{\infty}$ based on a Markov Chain where u_k is drawn from the distribution ρ_k at step k, and we want the following properties:

- the Markov Chain is invariant with respect to ρ
- the total variation distance between ρ_k and the target distribution ρ converges to zero (*ergodicity*):
$$\lim_{k\to\infty} d_{TV}(\rho_k, \rho) = 0.$$
- the empirical distribution generated by the $\{u_k\}$ converges to ρ; more precisely, for some class of test functions φ, the average of $\varphi(u_k)$ with respect to K steps of the Markov chain converges to the expectation of φ over the target distribution (*sample path ergodicity*):
$$\lim_{k\to\infty} \frac{1}{K} \sum_{k=1}^{K} \varphi(u_k) = \rho(\varphi).$$

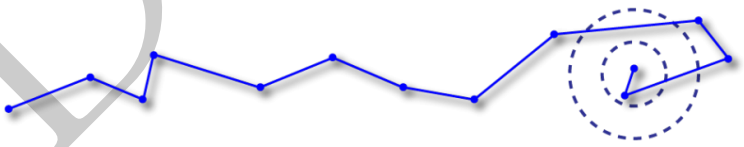


Figure 10 The Markov Chain samples points from distribution ρ_k at step k, and the sampling distribution converges towards the target distribution ρ whose high density regions are represented by the dashed circles.

The idea is illustrated in Figure 10: after an initial number of *burn-in* steps the samples from the chain start to concentrate in regions where the target distribution has greatest weight. To address the design and analysis of MCMC methods in great

[1] The MCMC methodology is readily extended to be used more generally for approximating certain multi-dimensional integrals whenever they may be rewritten as an expectation.

generality and depth is beyond the scope of a single chapter; an entire book could be devoted to this subject. We will instead focus on a particular class of MCMC methods, known collectively as Metropolis-Hastings Algorithms. And instead of studying the three desired properties in detail, we will focus on proving invariance for general Metropolis-Hastings methods, and on a specific choice of such a method for which it is possible to prove exponential convergence of ρ_k to ρ in the total variation norm – *geometric ergodicity*. We will not discuss sample path ergodicity, noting simply that a general abstract methodology exists to deduce it from geometric ergodicity.

7.2 The Metropolis-Hastings Algorithm

Here we outline the Metropolis-Hastings algorithm. The algorithm has two ingredients: a *proposal distribution* $q(u, v)$, which is a Markov transition kernel; and an acceptance probability $a(u, v)$ which will be used to convert the given Markov transition kernel into a kernel $p(u, v)$ which is invariant with respect to the given target $\rho(u)$. Given that we have the k^{th} iterate of the Markov chain u_k we generate u_{k+1} by drawing v_* from the distribution $q(u_k, \cdot)$ and accepting the result, which means setting $u_{k+1} = v_*$, with probability $a(u_k, v)$, or instead setting $u_{k+1} = u_k$ with the remaining probability $1 - a(u_k, v)$. The acceptance probability is given by

$$a(u, v) = \min\left(\frac{\rho(v)q(v, u)}{\rho(u)q(u, v)}, 1\right) \tag{7.1}$$

Algorithm 1 gives the precise definition of the the steps just described:

Algorithm 7.1 Metropolis-Hastings Algorithm

1: **Input**: Target distribution $\rho(u)$, Markov kernel $q(u, v)$, number of samples L

2: **Initial Draw**: Draw initial sample u_1.

3: **Subsequent Samples**: For $k = 1, 2, \ldots, L$, perform the following steps:

1. Draw $v_\star \sim q(u_k, \cdot)$
2. Calculate the acceptance probability $a_k := a(u_k, v_\star)$.
3. Update

$$u_{k+1} = \left\{ \begin{array}{ll} v_\star, & \text{w.p. } a_k \\ u_k, & \text{otherwise} \end{array} \right\}$$

4: **Output** Samples $u_1, u_2, \ldots, u_L$

The accept-reject step may be implemented by drawing, independently from the proposal, a uniformly distributed random variable r_k in the interval $[0, 1]$. If $a_k \in [0, r_k)$ then the proposal is accepted ($u_{k+1} = v_*$); it is rejected ($u_{k+1} = u_k$) otherwise.

7.3 Invariance of the Target Distribution ρ

To show invariance of the target distribution ρ, we utilize the idea of detailed balance. Metropolis-Hastings (and other MCMC methods such as Gibbs sampling) rely on this idea, which we detail below.

7.3.1 Detailed Balance and its Implication

A Markov Chain exhibits *detailed balance with respect to probability measure* ρ if the Markov kernel $p(u, v)$ defining the algorithm satisfies:

$$\rho(u)p(u, v) = \rho(v)p(v, u) \qquad \forall u, v \in \mathbb{R}^n$$

Corollary 7.1. *If Markov kernel p satisfies detailed balance with respect to the distribution ρ, then ρ is invariant for the Markov kernel.*

Proof. Integrating the expression for detailed balance with respect to u, and using that $p(\cdot, \cdot)$ is a Markov kernel so that $p(v, \cdot)$ integrates to 1 for every v, we obtain

$$\int_{\mathbb{R}^N} \rho(u)p(u, v)du = \int_{\mathbb{R}^N} \rho(v)p(v, u)du = \rho(v).$$

Assume that u is distributed according to pdf ρ. Then the expression on the left-hand side of the preceding identity represents the probability of being in state v after taking one step of the Markov chain, since $p(u, v)$ represents the probability of transitioning from u to v. Since the right-hand side of the identity is $\rho(v)$ we see that the distribution ρ is invariant under Markov kernel p. □

Another perspective from which to look at detailed balance is through the flow diagram shown in Figure 11. The flow to/from any node must be the same under detailed balance, so the density at each node must remain constant at each step, as shown in Figure 11(b); Figure 11(a) fails to satisfy this.

7.3.2 Detailed Balance and the Metropolis-Hastings Algorithm

Consider a generic Markov kernel $q(u, v)$ as shown in Figure 11(a), which does not satisfy detailed balance, which we accept-reject according to the Metropolis algorithm. Intuitively, this Metropolis-Hastings kernel, limits the "flow" between nodes such that detailed balance is satisfied with respect to the distribution ρ, as shown in Figure 11(b). There the red-highlighted numbers represent where the transition probabilities of the kernel have been reduced. We have the following theorem proving detailed balance of the Metropolis Hastings kernel:

Theorem 7.2 (Metropolis-Hastings and Detailed Balance)**.** *The Metropolis-Hastings kernel satisfies detailed balance with respect to the distribution ρ.*

Proof. We consider two cases:

1 $p(u_k, u_{k+1})$ where $u_{k+1} \neq u_k$

2 $p(u_k, u_k)$ where $u_{k+1} = u_k$.

We now establish that, in both Cases 1 and 2, detailed balance is satisfied under the Metropolis-Hastings kernel p, comprising the required proof.
Case 1: In this case we have necessarily accepted the proposed move and so

$$\begin{aligned} p(u_k, u_{k+1}) &= \min\left(1, \frac{\rho(u_{k+1})q(u_{k+1}, u_k)}{\rho(u_k)q(u_k, u_{k+1})}\right) q(u_k, u_{k+1}) \\ &= \frac{1}{\rho(u_k)} \times \min\left(\rho(u_k)q(u_k, u_{k+1}), \rho(u_{k+1})q(u_{k+1}, u_k)\right). \end{aligned}$$

Thus, invoking symmetry,

$$\rho(u_k)p(u_k, u_{k+1}) = \min\left(\rho(u_k)q(u_k, u_{k+1}), \rho(u_{k+1})q(u_{k+1}, u_k)\right) = \rho(u_{k+1})p(u_{k+1}, u_k).$$

Case 2: In this case we may arrive back at u_k either by having proposed a move there, and accepting it, or by proposing a move away from u_k, and rejecting it; thus we have

$$p(u_k, u_k) = a(u_k, u_k)q(u_k, u_k) + \Big(1 - \int_{u \neq u_k} q(u_k, u)a(u_k, u)du\Big).$$

Since $u_{k+1} = u_k$, it is straightforward that $\rho(u_k)p(u_k, u_{k+1}) = \rho(u_{k+1})p(u_{k+1}, u_k)$.

□

7.4 Convergence to the Target Distribution

The Metropolis-Hastings algorithm is a method to modify an arbitrary Markov kernel, through accept-reject, in such a way that the target measure ρ is invariant. If we initialize this chain with distribution ρ, it will therefore keep producing samples from the invariant distribution. But our original problem was exactly that we were not able to sample from ρ. Therefore, for the Markov chain to be of value, we need to ensure that even if we initialize the chain by drawing u_0 from another, tractable, initial distribution ρ_0, we eventually *converge* to ρ. This is known as *ergodicity*. Ergodicity does not need to be true in general, as illustrated by the chain depicted in Figure 12.

We will now identify conditions under which a Markov chain is ergodic, that is that it converges to its invariant distribution, for any initial distribution. We first consider a chain with a finite statespace, to understand the mechanisms behind ergodicity, and then study a specific Metropolis-Hastings method, known as the *preconditioned Crank-Nicolson (pCN)* method, which applies to targets ρ defined by their density with respect to a Gaussian distribution.

7.4.1 Finite State Space

We consider a Markov chain on the finite state space $S = \{1, \cdots, N\}$. We illustrate a *coupling* approach to proving ergodicity and then, in the next subsection, use it to study the pCN method on a continuous state space.

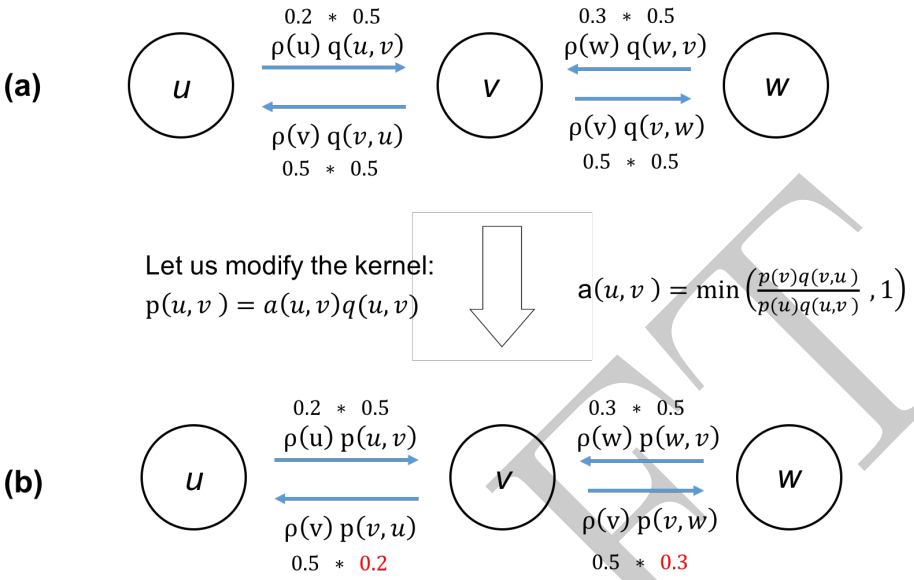


Figure 11 Toy representation of MCMC with 3 states (u, v, w). The numbers associated with each term represent an example case. (a) The top chain represents a Markov Chain with Markov kernel q which does not satisfy detailed balance with respect to ρ. (b) The bottom chain utilizes Markov kernel p, which satisfies detailed balance by modifying the transition probabilities of q.

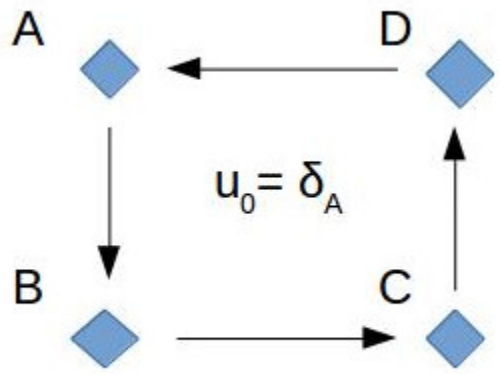


Figure 12 The arrows represent transitions with probability one in a four state Marko chain. The invariant distribution is the uniform distribution but for $\rho_0 = \delta_A$, we do not have ergodicity.

Theorem 7.3 (Ergodicity in Finite State Spaces)**.** *Let u_k be Markov chain with state space S, transition kernel p, and initial distribution ρ_0. Assume that*

$$\varepsilon := \min_{i,j \in S} p(i,j) > 0. \tag{7.2}$$

Then, p has a unique invariant distribution ρ, and the following convergence result holds:

$$d(\rho_k, \rho)_{TV} \leq (1-\varepsilon)^k.$$

Proof. First note that the markov kernel p maps a probability distribution on S into another probability distribution on S; it thus maps a convex set into itself. By Brouwer's fixed point theorem it follows that p has a fixed point in this space, ensuring that a stationary distribution exists, and we denote it by ρ, a probability vector on S. Proving convergence to equilibrium amounts to "forgetting the past", to show that the long time behavior of the Markov chain does not depend on the initial distribution ρ_0 and in fact converges to ρ.

In general, u_{k+1} will be strongly dependent on u_k, but the condition given in (7.2) implies that there is always some residual chance that the chain jumps to any new state, at each step, independently of where it is currently located, u_k. We will show that this residual probability of the chain to make a "totally random" move will diminish the stochastic dependence on u_0 as k increases. To formalize this idea, let Z_k be i.i.d. Bernoulli random variables with $\mathbb{P}[Z_k = 1] = \varepsilon$ and $\mathbb{P}[Z_k = 0] = 1 - \varepsilon$; furthermore assume that the sequence $\{Z_k\}$ is independent of the randomness defning draws from $\{p(u_k, \cdot)\}$. Because of the lower bound on p we may define a new Markov chain as follows:

$$u_{k+1} \sim \begin{cases} s(u_k, \cdot), & \text{for } Z_k = 0 \\ r(u_k, \cdot), & \text{for } Z_k = 1, \end{cases}$$

where we define

$$s(i,j) :- \frac{p(i,j) - \varepsilon r(i,j)}{1-\varepsilon}.$$

and where r is the uniform transition kernel with equal probability of transitioning to each state in $S : r(i,j) = N^{-1}$ for all $(i,j) \in S \times S$. We may now compute

$$\begin{aligned} \mathbb{P}[u_{k+1} = j | u_k = i] &= \varepsilon\, \mathbb{P}[u_{k+1} = j | u_k = i, Z_k = 1] + (1-\varepsilon)\, \mathbb{P}[u_{k+1} = j | u_k = i, Z_k = 0] \\ &= \varepsilon r(i,j) + p(i,j) - \varepsilon r(i,j) \\ &= p(i,j). \end{aligned}$$

Thus the new transition rule is equivalent in law to the one given by p. However, by introducing the ancillary random variables Z_k, we have made explicit the concept of "forgetting the past entirely, with a small probability" at every step. We may now use this to complete the proof. Define an arbitrary test function $\varphi : S \mapsto \mathbb{R}$, with $|\varphi| \leq 1$

and $Y \coloneqq \min(k \in \mathbb{N} : Z_k = 1)$. Then

$$\begin{aligned}\mathbb{E}[\varphi(u_k)] &= \mathbb{E}[\varphi(u_k)|Y \geq k]\,\mathbb{P}[Y \geq k] + \sum_{l=0}^{k-1} \mathbb{E}[\varphi(u_k)|Y = l]\,\mathbb{P}[Y = l] \\ &= \underbrace{\mathbb{E}[\varphi(u_k)|Y \geq k]\,\mathbb{P}[Y \geq k]}_{|\cdot| \leq (1-\varepsilon)^k} + \underbrace{\sum_{l=0}^{k-1} \mathbb{E}_{u_0 \sim \mathfrak{u}(\cdot)}[\varphi(u_{k-l})]\,\mathbb{P}[Y = l]}_{\text{independent of original initial distribution}} \quad ,\end{aligned}$$

where $\mathfrak{u}$ denotes the uniform distribution on S.

Now consider two Markov chains, one initialized by sampling from an arbitrary distribution ρ_0 on S, and the other by sampling from the desired target ρ. We denote the distribution at the k^{th} step of these Markov chains by ρ_k and ρ'_k respectively. By employing the preceding identity in the definition of the total variation distance, noting that the contribution which is independent of the initial distribution will cancel in the two different Markov chains, we obtain[2]

$$d(\rho_k, \rho'_k)_{\mathtt{TV}} = \frac{1}{2} \sup_{|\varphi| \leq 1} \left| \mathbb{E}^{\rho_k}[\varphi(u)] - \mathbb{E}^{\rho'_k}[\varphi(u)] \right| \leq (1-\varepsilon)^k.$$

Since ρ'_k is from the Markov chain initialized to ρ, and ρ is the invariant distribution of the Markov chain, we have that $\rho'_k = \rho$. The desired result follows. □

Before extending the above argument to a setting with continuous state space, we make two remarks:

Remark 7.4. The coupling proof we have just exhibited may be generalized in a number of ways; in particular:

- The distribution r need not be uniform; it was only chosen so for convenience. What is important is that $r_{i,j}$ is lower bounded, independently of i, for all j. Adapting r to the p at hand, might in some cases greatly improve the above bound – a larger ϵ might be identified.
- Convergence to equilibrium can also be shown if condition (7.2) holds with p replaced by the k-step transition kernel p^k. Again, for some chains this may yield faster bounds on the convergence to equilibrium.

□

7.4.2 The pCN Method

The coupling argument used in the previous subsection, for a finite and discrete state space, may also be employed to study ergodicity of Markov chains on a continuous statespace. To illustrate this we consider a particular Metropolis-Hastings algorithm,

[2] The characterization of total variation distance used here reduces to the simpler "half the L^1 norm of the densities", introduced in chapter 3, for densities which are not Dirac's.

the pCN method, applied to a specific Bayesian inverse problem. Before we get into the details of the inverse problem we describe the idea behind the pCN method at a high level. The idea is this. If desired target distribtion has the form $\rho(v) = \rho_0(v) \times \ell(v)$ for some other distribution ρ_0, and if the Metropolis-Hastings proposal q satisfies detailed balance with respect to ρ_0, then (7.1) simplifies to give

$$a(u,v) = \min\left(\frac{\rho_0(v)\ell(v)q(v,u)}{\rho(u)\ell(u)q(u,v)}, 1\right) = \min\left(\frac{\ell(v)}{\ell(u)}, 1\right). \tag{7.3}$$

In many inverse problems the prior is sufficiently simple that constructing proposals which satisfy detailed balance with respect to it is straightforward; Gaussian distributions provide a natural family of examples. Formula (7.3) shows that then the acceptance probability has a very intuitive form in terms of the likelihood ℓ. In particular proposals are always accepted if they increase the likelihood.

The inverse problem we study has likelihood and prior as follows:

Assumption 7.5. *We make the following assumptions concerning our Bayesian inverse problem:*

- *The likelihood has the form*

$$\pi(y - G(u)) \propto \exp(-l(u,y)),$$

  where the loss function fulfills $0 < l^- \leq l(u,y) \leq l^+ < \infty$, $\forall (u,y) \in \mathbb{R}^N \times \mathbb{R}^J$.

- *The prior density* $\rho_0 \propto g(u)\mathbb{1}_B(u)$, *for a bounded set* $B \subset \mathbb{R}^N$ *and* g *the Gaussian density* $N(0, C_0)$.

Under Assumption 7.5 we obtain for the posterior density ρ:

$$\rho(u) \propto \exp(-l(u,y))g(u)\mathbb{1}_B(u).$$

In the notation of the preceding discussion we will take

$$\rho_0(u) = g(u), \quad \ell(u)\exp(-l(u,y))\mathbb{1}_B(u).$$

The pCN method has proposal kernel which is invariant with respect to the Gaussian g and is defined via:

$$v_* \sim \left(1 - \beta^2\right)^{1/2} u_k + \beta\xi_k, \quad \xi_k \sim N(0, C_0). \tag{7.4}$$

We let ρ_k denote the Markov chain that results from a Metropolis-Hasting accept-reject criterion applied to this proposal kernel, with u_0 initially distributed according to an arbitrary density ρ_0.

Theorem 7.6 (Ergodicity for pCN Method)**.** *Assume that we apply the pCN method to sample from posterior density* ρ *arising from Assumptions 7.5 with initial condition drawn from any density supported on* B. *Then there exists a constant* $\varepsilon \in (0,1)$ *such that*

$$d_{\mathrm{TV}}(\rho_k, \rho) \leq (1 - \varepsilon)^k,$$

where ρ_k *is the law of the pCN Metropolis-Hastings algorithm.*

Recall the notation for the covariance weighted inner-product and resulting norm described in the introduction to these notes. Before proving the above theorem, we will prove one auxiliary lemma.

Lemma 7.7. *Let the assumptions of Theorem 7.6 hold. The quantity*

$$z := g(v) \exp^{-\frac{1}{2\beta^2}\left|u-\left(1-\beta^2\right)^{1/2}v\right|^2_{C_0}}$$

is symmetric in u and v. As a consequence the proposal (7.4) defines a kernel which satisfies detailed balance with respect to g and we have the following expression for the pCN accept-reject probability:

$$a(u,v) = \min\{1, \exp(l(u,y) - l(v,y))\mathbb{1}_B(v)\}.$$

Proof. We see that, since $u_0 \in B$, we have $u_k \in B$ for all k as any proposed move out of B will be rejected. Thus $\mathbb{1}_B(u)$ is dropped from the formula for the acceptance probability $a(u,v)$ and the formula becomes

$$a(u,v) := \min\left\{1, \frac{\exp(-l(v,y))g(v)\mathbb{1}_B(v)\exp\left(-\frac{1}{2\beta^2}\left|u-\left(1-\beta^2\right)^{1/2}v\right|^2_{C_0}\right)}{\exp(-l(u,y))g(u)\exp\left(-\frac{1}{2\beta^2}\left|v-\left(1-\beta^2\right)^{1/2}u\right|^2_{C_0}\right)}\right\}.$$

Thus the stated symmetry of z results in the given expression for the accept-reject probability, and so it remains to prove that symmetry. This follows by noting that, ignoring an additive constant independent of u and v,

$$\begin{aligned}
-\ln z &= \frac{1}{2}|v|^2_{C_0} + \frac{1}{2\beta^2}\left|u-\left(1-\beta^2\right)^{1/2}v\right|^2_{C_0} \\
&= \left(\frac{1}{2} + \frac{\left(1-\beta^2\right)}{2\beta^2}\right)|v|^2_{C_0} + \frac{1}{2\beta^2}|u|^2_{C_0} - \frac{\left(1-\beta^2\right)^{1/2}}{\beta^2}\langle u,v\rangle_{C_0} \\
&= \frac{1}{2\beta^2}\left(|v|^2_{C_0} + |u|^2_{C_0}\right) - \frac{\left(1-\beta^2\right)^{1/2}}{\beta^2}\langle u,v\rangle_{C_0}.
\end{aligned}$$

Noting that $z = z(u,v) = g(v)q(v,u)$ where q is the proposal defined by (7.4) also establishes the stated detailed balance. □

Using this lemma, we can prove ergodicity much as we did in the previous subsection in the finite state space setting. The main idea is that, restricted to the bounded set B, the probability density of the transition kernel will be bounded away from zero by some ε. Splitting off a "forgetful part" that is triggered with probability ε will then yield the result.

Proof of Theorem 7.6. We see that, since $u_0 \in B$, we have $u_k \in B$, for all k. We use the notation $\mathbb{P}_k[\cdot] := p(u_k,\cdot)$. Note that $\mathbb{P}_k[u_{k+1} \in A]$, for fixed but arbitrary measurable set A, is equal to

$$\mathbb{P}_k[u_{k+1} \in A | u_{k+1} = v_*]\,\mathbb{P}_k[u_{k+1} = v_*] + \mathbb{P}_k[u_{k+1} \in A | u_{k+1} = u_k]\,\mathbb{P}_k[u_{k+1} = u_k]$$

and so may be lower bounded by

$$\mathbb{P}_k[v_* \in A]a(u_k, v_*) \geq \varepsilon r(A).$$

Here r is the Gaussian measure $N\Big(0, \frac{\beta^2}{2}C_0\Big)$ and

$$\varepsilon = 2^{-N/2} \exp\left(\frac{-(1-\beta^2)}{\beta^2} \sup_{u\in B} |u|^2_{C_0}\right) \exp\Big(l^- - l^+\Big).$$

This follows because the acceptance probability is bounded below by

$$a(u_k, v_*) \geq \exp(l(u_k, y) - l(v_*, y)) \geq \exp(l^- - l^+)$$

and because

$$\begin{aligned}\mathbb{P}_k[v_* \in A] &= \int_A \frac{1}{(2\pi\beta^2)^{N/2}(\det C_0)^{\frac{1}{2}}} \exp\Big(-\frac{1}{2\beta^2}\Big|v - (1-\beta^2)^{\frac{1}{2}}u_k\Big|^2_{C_0}\Big)dv \\ &\geq \int_A \frac{1}{2^{N/2}(\pi\beta^2)^{N/2}(\det C_0)^{\frac{1}{2}}} \exp\Big(-\frac{1}{\beta^2}|v|^2_{C_0} - \frac{(1-\beta^2)}{\beta^2}|u_k|^2_{C_0}\Big)dv.\end{aligned}$$

Analogously to the discrete proof, we now define Z_k to be i.i.d. Bernoulli random variables with $\mathbb{P}[Z_k = 1] = \varepsilon$, independently of all other randomness in what follows, and consider the transition rule

$$u_{k+1}|u_k \sim \begin{cases} s(u_k, \cdot), & \text{for } Z_k = 0 \\ r(\cdot), & \text{for } Z_k = 1, \end{cases}$$

where we define

$$s(u, A) := \frac{p(u, A) - \varepsilon r(A)}{1 - \varepsilon}.$$

Just as in the discrete case, one can check that the resulting Markov chain is equal in law to that generated by kernel $p(\cdot, \cdot)$. Exponential convergence is then concluded in exactly the same way as in the discrete case. □

7.5 Discussion and Bibliography

The book [38] is a useful basic introduction to MCMC and the book [16] presents state of the art as of 2010. The paper [22] overviews the pCN method, and related MCMC algorithms specifically designed for inverse problems and other sampling problems in high dimensional state spaces. The book [71] describes the coupling method in a general setting. The book [82] contains a wide-ranging presentation of Markov chains, and their long-time behavior, including ergodicity, and coupling. Furthermore the book describes the general methodology for going from convergence of expectations in (possibly weighted) total variation metrics to sample path ergodicity and almost sure convergence of time averages, a topic we did not cover in this chapter. The paper [81] describes the coupling methodology in the context of stochastic differential equations and their approximations.

8 The Filtering Problem and Well-Posedness

In this chapter we introduce data assimilation problems in which the model of interest, and the data associated with it, have a time-ordered nature. We distinguish between the filtering problem (on-line) in which the data is incorporated sequentially as it comes in, and the smoothing problem (off-line) which is a specific instance of the inverse problems that have been the subject of the preceding chapters.

8.1 Formulation of Filtering and Smoothing Problems

Consider the *stochastic dynamics model* given by:

$$\begin{aligned} v_{j+1} &= \Psi(v_j) + \xi_j,\ j \in \mathbb{Z}^+ \\ v_0 &\sim N(m_0, C_0),\ \xi_j \sim N(0, \Sigma) \text{ i.i.d.}, \end{aligned}$$

where we assume that v_0 is independent of the sequence $\{\xi_j\}$; this is often written as $v_0 \perp \{\xi_j\}$. Now we add the *data model* given by:

$$\begin{aligned} y_{j+1} &= h(v_{j+1}) + \eta_{j+1},\ j \in \mathbb{Z}^+ \\ \eta_j &\sim N(0, \Gamma) \text{ i.i.d.}, \end{aligned}$$

where we assume that $\{\eta_j\} \perp v_0$ for all j and $\{\eta_k\} \perp \{\xi_j\}$. for all j, k. We make the following assumptions:

Assumption 8.1. *Both Σ and Γ are positive-definite symmetric. Further, we have $\Psi \in C(\mathbb{R}^n, \mathbb{R}^n)$ and $h \in C(\mathbb{R}^n, \mathbb{R}^m)$.*

We define the following sequences of states as follows:

$$v = \{v_0, ..., v_J\},\ y = \{y_1, ..., y_J\}, : \text{and } Y_j = \{y_1, ..., y_j\}$$

The sequence v is often termed the *signal* and the sequence y the *data*.

Definition 8.2 (The Smoothing Problem)**.** The *smoothing problem* is to find the probability density $p(v_0, \cdots, v_J) := \mathbb{P}(v|y) = \mathbb{P}(\{v_0, ..., v_J\}|\{y_1, ..., y_J\})$ on $\mathbb{R}^{(J+1)n}$ for some fixed integer J.

Definition 8.3 (The Filtering Problem)**.** The *filtering problem* is to find the probability densities $\rho_j(v_j) := \mathbb{P}(v_j|Y_j)$ on $\mathbb{R}^n$ for $j = 1, \ldots, J$.

Remark 8.4. We note the following identity:

$$\int p(v_0, ..., v_J) dv_0 dv_1 ... dv_{J-1} = \rho_J(v_J)$$

This expresses the fact that the marginal of the smoothing distribution at time J corresponds to the filtering distribution at time J. Note also that, in general, for $j < J$

$$\int p(v_0, ..., v_j) dv_0 ... dv_{j-1} dv_j dv_{j+1} ... dv_J \neq \rho(v_j),$$

since the expression on the left-hand side of the equation depends on data Y_J, whereas that on the right-hand side depends only on Y_j, and $j < J$.

8.2 The Smoothing Problem

8.2.1 Formula for pdf of the Smoothing Problem

The prior is the probability distribution on v implied by the dynamics model; the posterior conditions this on the data y from the data model. The smoothing distribution is the posterior distribution on $v|y$, found by combining the prior and the likelihood function. We now derive the prior and the likelihood separately.

The prior distribution can be derived as follows:

$$\begin{aligned}\mathbb{P}(v) &= \mathbb{P}(v_J, v_{J-1}, ..., v_0) \\ &= \mathbb{P}(v_J|v_{J-1}, ..., v_0)\mathbb{P}(v_{J-1}, ..., v_0) \\ &= \mathbb{P}(v_J|v_{J-1})\mathbb{P}(v_{J-1}, ..., v_0)\end{aligned}$$

The third equality comes from the Markov, or memoryless, property which follows from the independence of the elements of the sequence $\{\xi_j\}$. By induction, we have:

$$\begin{aligned}\mathbb{P}(v) &= \prod_{j=0}^{J-1} \mathbb{P}(v_{j+1}|v_j)\mathbb{P}(v_0) \\ &\propto \exp(-r(v)) \\ &= \frac{1}{Z_0}\exp(-r(v)) \\ &=: p_0(v),\end{aligned}$$

where

$$r(v) = \frac{1}{2}|v_0 - m_0|^2_{C_0} + \sum_{j=0}^{J-1}\frac{1}{2}|v_{j+1} - \Psi(v_j)|^2_{\Sigma}.$$

The likelihood function, which incorporates the measurements gathered from observing the system, depends only on the measurement model and may be derived as follows:

$$\begin{aligned}\mathbb{P}(y|v) &= \prod_{j=0}^{J-1} \mathbb{P}(y_{j+1}|v_0, ..., v_J) \\ &= \prod_{j=0}^{J-1} \mathbb{P}(y_{j+1}|v_{j+1}) \\ &\propto \exp(-l(v; y)),\end{aligned}$$

where

$$l(v; y) = \sum_{j=0}^{J-1}\frac{1}{2}|y_{j+1} - h(v_{j+1})|^2_{\Gamma}.$$

The factorization of $\mathbb{P}(y|v)$ in terms of the product of the $\mathbb{P}(y_{j+1}|v_{j+1})$ follows from the independence of the elements of $\{\eta_j\}$ and the fact that the observation at time $j+1$ depends only on the state at time $j+1$.

Using Bayes Theorem 1.2, the posterior distribution for the smoothing problem becomes:

$$\begin{aligned} p(V) &\propto \mathbb{P}(y|v)p_0(v) \\ &= \frac{1}{Z}\exp(-r(v) - l(v;y)). \end{aligned}$$

Note that $v \in \mathbb{R}^{n(J+1)}$, and $y \in \mathbb{R}^{m(J)}$. We let $N = n(J+1)$ and $M = m(J)$.

8.2.2 Well-Posedness of the Smoothing Problem

Now we study the well-posedness of the posterior distribution with respect to perturbations in the data. To this end we consider two posterior distributions corresponding to different observed data sequences y, y':

$$\begin{aligned} p(v) &:= \mathbb{P}(v|y) = \frac{1}{Z}\exp(-r(v) - l(v;y)) \\ p'(v) &:= \mathbb{P}(v|y') = \frac{1}{Z}\exp(-r(v) - l(v;y')). \end{aligned}$$

For the well-posedness proof, we make the following assumptions:

Assumption 8.5. *There is finite non-negative constant R such that the data y, y' and the observation function h satisfy the following bounds:*

- $y, y' \in B(0,R) \subset \mathbb{R}^M$;
- $g(v) = \sum_{j=1}^{J}(|y_j|^2 + |y_j'|^2 + |h(v_j)|^2)$, *and* $\mathbb{E}^{p_0} g(v) < R$.

The ball of radius R in $\mathbb{R}^M$ is with respect to the Euclidean norm, and p_0 is the prior on v derived above.

Theorem 8.6 (Well-posedness of Smoothing)**.** *Under Assumption 8.5, $\exists \kappa = \kappa(z) \in [0,\infty)$ such that*

$$d_H(p, p') \leq \kappa |y - y'|.$$

Proof. Throughout the proof κ may change from instance to instance but is always dependent only on Z. Furthermore all integrals are over $\mathbb{R}^N$. Let z, z' and z_0 denote the normalization constants for the posteriors corresponding to data y, y' and the prior respectively. Then

$$\begin{aligned} \frac{z}{z_0} &= \frac{\int e^{-l(v;y)-r(v)}dv}{\int e^{-r(v)}dv} \\ &= \int e^{-l(v;y)} p_0(v)dv \\ &\geq \int e^{-g} p_0(v)dv \\ &\geq e^{-R}\mathbb{P}^{p_0}(|g| < R) \\ &\geq e^{-R}(1 - \mathbb{P}^{p_0}(|g| \geq R)) \\ &\geq e^{-R}(1 - \frac{\mathbb{E}^{p_0}|g|}{R}) \\ &> 0, \end{aligned}$$

for R large enough. Note that we used:

$$\exp(-r(v))dv = z_0 p_0(v)dv \tag{8.3}$$

and the Markov inequality. Using a similar argument, we can also show that z_0 is positive using:

- $\Psi(.)$ is bounded on the set $\{|v| \leq R\}$.
- The Gaussian measure $N(m_0, C_0)$ is strictly positive on the set $\{|v| \leq R\}$.

Thus $z_0, z, z' > 0$.

Now note that, using (8.3),

$$\begin{aligned} d_{\mathtt{H}}(p,p')^2 &= \frac{1}{2}\int |\sqrt{p(v)} - \sqrt{p'(v)}|^2 dv \\ &\leq I_1 + I_2, \end{aligned}$$

where

$$I_1 = \frac{z_0}{2}\int \frac{1}{z}|e^{-\frac{1}{2}l(v;y)} - e^{-\frac{1}{2}l(v;y')}|^2 p_0(v)dv,$$

and

$$\begin{aligned} I_2 &= \frac{z_0}{2}\Big|\frac{1}{\sqrt{z}} - \frac{1}{\sqrt{z'}}\Big|^2 \int e^{-l(v;y')}p_0(v)dv \\ &= z'\Big|\frac{1}{\sqrt{z}} - \frac{1}{\sqrt{z'}}\Big|^2 \\ &= \frac{1}{z}\Big|\sqrt{z} - \sqrt{z'}\Big|^2 \\ &= \frac{1}{z}\Big(\frac{|z-z'|}{|\sqrt{z}+\sqrt{z'}|}\Big)^2 \\ &\leq \kappa|z-z'|^2. \end{aligned}$$

Now since e^{-x} is Lipschitz-1, we have that ,

$$\begin{aligned} |z-z'| &\leq z_0 \int |e^{-l(v;y)} - e^{-l(v;y')}p_0(v)|dv \\ &\leq z_0 \int |l(v;y) - l(v;y')|p_0(v)dv. \end{aligned} \tag{8.4}$$

But we also know by some algebraic manipulations and Cauchy-Schwarz:

$$\begin{aligned} |l(v;y) - l(v;y')| &\leq \frac{1}{2}\sum_{j=0}^{J-1} |y_{j+1} - y'_{j+1}|_\Gamma |y_{j+1} + y'_{j+1} - 2h(v_{j+1})|_\Gamma \\ &\leq \kappa|y - y'|g^{\frac{1}{2}}. \end{aligned}$$

Putting this together, we have:

$$|z - z'| \leq z_0 \kappa |y - y'| \int g^{\frac{1}{2}} p_0(v) dv$$

Now $\mathbb{E}^{p_0} g^{\frac{1}{2}} \leq (\mathbb{E}^{p_0} g)^{\frac{1}{2}} (\mathbb{E}^{p_0} 1)^{\frac{1}{2}} = (\mathbb{E}^{p_0} g)^{\frac{1}{2}}$. Thus, we have shown that $|z - z'| \leq \kappa |y - y'|$ from which it follows that $I_2 \leq \kappa |y - y'|^2$.

Now we bound I_1:

$$\begin{aligned} I_1 &\leq \kappa \int |l(v; y) - l(v; y')|^2 p_0(v) dv \\ &\leq \kappa |y - y'|^2 \int g p_0(v) dv \\ &\leq \kappa |y - y'|^2. \end{aligned}$$

Combining the bounds on I_1 and I_2 gives the desired result. □

8.3 The Filtering Problem

8.3.1 Formula for pdf of the Filtering Problem

The key conceptual issue to appreciate concerning the filtering problem, in comparison with the smoothing problem, is that interest is focused on solving for, or approximating, a sequence of probably distributions, defined in an iterative fashion as the data is acquired sequentially. This rests on the formula

$$\rho_{j+1} = L_j P \rho_j,$$

where P is a *linear* Markov map and L_j is a *nonlinear* likelihood map (Bayes' Theorem). We introduce $\widehat{\rho}_{j+1} = \mathbb{P}(v_{j+1}|Y_j)$ and recall that $\rho_j = \mathbb{P}(v_j|Y_j)$. Then the maps P and L_j are defined by

$$\begin{aligned} \textbf{Prediction Step:} \quad & \widehat{\rho}_{j+1} = P \rho_j \\ \textbf{Analysis Step:} \quad & \rho_{j+1} = L_j \, \widehat{\rho}_j \end{aligned} \tag{8.5}$$

The map P is sometimes termed *prediction*; the map L_j is termed *analysis*.

Throughout this section, all integrals are over $\mathbb{R}^N$. First we derive the linear Markov map P. By the Markov property of the stochastic dynamics model we have

$$\begin{aligned} \widehat{\rho}_{j+1}(v_{j+1}) &= \mathbb{P}(v_{j+1}|Y_j) \\ &= \int \mathbb{P}(v_{j+1}|Y_j, v_j) \mathbb{P}(v_j|Y_j) dv_j \\ &= \int \mathbb{P}(v_{j+1}|v_j) \mathbb{P}(v_j|Y_j) dv_j \\ &= \int \mathbb{P}(v_{j+1}|v_j) \rho_j(v_j) dv_j \\ &= \frac{1}{((2\pi)^n \det \Sigma)^{\frac{1}{2}}} \int \exp\Big(-\frac{1}{2} |v_{j+1} - \Psi(v_j)|^2_\Gamma\Big) \rho_j(v_j) dv_j. \end{aligned} \tag{8.6}$$

This defines the linear integral operator P.

Now we derive the nonlinear likelihood map L_j by using Bayes' Theorem:

$$\rho_{j+1} = L_j \widehat{\rho}_{j+1}.$$

We note that

$$\begin{aligned}
\rho_{j+1}(v_{j+1}) &= \mathbb{P}(v_{j+1}|Y_{j+1}) \\
&= \mathbb{P}(v_{j+1}|Y_j, y_{j+1}) \\
&= \frac{\mathbb{P}(y_{j+1}|v_{j+1}, Y_j)\mathbb{P}(v_{j+1}|Y_j)}{\mathbb{P}(y_{j+1}|Y_j)} \\
&= \frac{\mathbb{P}(y_{j+1}|v_{j+1})\mathbb{P}(v_{j+1}|Y_j)}{\mathbb{P}(y_{j+1}|Y_j)} \\
&= \frac{\exp(-\frac{1}{2}|y_{j+1} - h(v_{j+1})|^2_\Gamma)\widehat{\rho}_{j+1}(v_{j+1})}{\int \exp(-\frac{1}{2}|y_{j+1} - h(v_{j+1})|^2_\Gamma)\widehat{\rho}_{j+1}(v_{j+1})dv_{j+1}}.
\end{aligned} \tag{8.7}$$

This defines the nonlinear map L_{j+1} through multiplication by the likelihood, and then normalization to a probability measure.

8.3.2 Well-Posedness of the Filtering Problem

Now, we would like to consider the well-posedness of the posterior distribution for the filtering problem. We consider the following two posterior distributions:

$$\begin{aligned}
\rho_J &= \mathbb{P}(v_j|y_j) = \frac{1}{Z}\exp(-r(v) - l(v; y)), \\
\rho'_J &= \mathbb{P}(v_j|y'_j) = \frac{1}{Z}\exp(-r(v) - l(v; y')).
\end{aligned}$$

We use the relationship between the smoothing problem and the filtering problem to establish the well-posedness of the filtering problem.

Corollary 8.7 (Well-posedness of Filtering)**.** *Under Assumption 8.5, there* $\exists \kappa = \kappa(R)$*, such that* $d_{TV}(\rho_J, \rho'_J) \leq \kappa|y - y'|$.

Proof. In this proof, we consider p, p' to be the posterior distributions from the smoothing problem $p = \mathbb{P}(v|y)$ and $p' = \mathbb{P}(v|y')$. We note that $\exists \kappa$ such that $d_{\mathrm{TV}}(p, p') \leq \kappa|y - y'|$ by Theorem 8.6 and by the fact that the Hellinger metric bounds the total variation metric (Lemma 3.2). Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $F : \mathbb{R}^{(n+1)J} \mapsto \mathbb{R}$. Then

$$\begin{aligned}
d_{\mathrm{TV}}(\rho_J, \rho_{J'}) &= \frac{1}{2}\sup_{|f|_\infty \leq 1} |\mathbb{E}^{\rho_J} f(v_J) - \mathbb{E}^{\rho_{J'}} f(v_J)| \\
&= \frac{1}{2}\sup_{|f|_\infty \leq 1} |\mathbb{E}^{p} f(v_J) - \mathbb{E}^{p'} f(v_J)| \\
&\leq \frac{1}{2}\sup_{|F|_\infty \leq 1} |\mathbb{E}^{p} F(v) - \mathbb{E}^{p'} F(v)| \\
&= d_{\mathrm{TV}}(p, p') \\
&\leq \kappa|y - y'|.
\end{aligned}$$

Here the first inequality follows from the fact that $\{|f|_\infty \leq 1\}$ can be viewed as a subset of $\{|F|_\infty \leq 1\}$. □

8.4 Optimization Perspective on Filtering Problems

In the subsequent sections, we introduce three popular filtering problems

8.5 Discussion and Bibliography

The book [70] gives a mathematical introduction to data assimilation; for further information on the smoothing problem as presented here, see section 2.3 in that book; for further information on the filtering problem as presented here, see section 2.4. The books [1, 95] give alternative foundational presentations of the subject. The books [59, 87, 78, 17] study data assimilation in the context of weather forecasting, oil reservoir simulation, turbulence modeling and geophysical sciences, respectively.

9 The Kalman Filter

In this chapter we study the filtering problem in the specific case where the state-transition and observation operators $\Psi(\cdot)$ and $h(\cdot)$ are both linear: the *Kalman filter*. The filtering distribution is then Gaussian, and entirely determined by the mean and covariance at each time. We start by recapping the filtering problem, then we find formulae for iterative updating of the Kalman filter mean and covariances as $j \mapsto j+1$. The remainder of the chapter then discusses the Kalman update formulae from an optimization point of view, followed by a result concerning the optimality of the Kalman filter.

9.1 Filtering Problem

Suppose we have a discrete-time dynamical system with noisy state transitions and noisy observations:

$$\begin{aligned}
\text{Dynamics Model:} \quad & v_{j+1} = \Psi(v_j) + \xi_j, \quad j \in \mathbb{Z}^+ \\
\text{Data Model:} \quad & y_{j+1} = h(v_{j+1}) + \eta_{j+1}, \quad j \in \mathbb{Z}^+ \\
\text{Probabilistic Structure:} \quad & v_0 \sim N(m_0, C_0), \quad \xi_j \sim N(0, \Sigma), \quad \eta_j \sim N(0, \Gamma) \\
\text{Probabilistic Structure:} \quad & v_0 \perp \{\xi_j\} \perp \{\eta_j\} \text{ independent}
\end{aligned}$$

$\Psi(\cdot)$ is the state-transition operator and $h(\cdot)$ is the observation operator. Additionally, recall that $Y_j := \{y_1, \ldots, y_j\}$ is the accumulated data up to time j. The *filtering problem* is to estimate the state at time j given the data from the past up to the present time j. That is, we want to determine the pdf $\rho_j = \mathbb{P}(v_j|Y_j)$. We also define $\widehat{\rho}_{j+1} = \mathbb{P}(v_{j+1}|Y_j)$ and recall that the evolution

$$\rho_{j+1} = L_j P \rho_j, \quad \rho_0 = N(m_0, C_0)$$

This can be rewritten as a prediction and analysis steps (8.5). Here P does not depend on j because the same Markov process governs the prediction step whereas L_j depends on j because at each step j the likelihood sees different data. P is a linear mapping, but L_j is nonlinear.

9.2 Kalman Filter

The Kalman filter is a special case of the the filtering problem when $\Psi(\cdot)$ and $h(\cdot)$ are linear maps. We make the following assumption:

Assumption 9.1.

$$\Psi(v) = Mv, \quad h(v) = Hv.$$

for matrices $M \in \mathbb{R}^{n\times n}$ and $H \in \mathbb{R}^{m\times n}$. Furthermore $m \leq n$.

The following theorem is a straightforward consequence of the linearity of Ψ and h:

Theorem 9.2. *(Gaussianity of Posterior Distributions) Under Assumption 9.1, ρ_0, $\{\rho_{j+1}\}_{j\in\mathbb{Z}^+}$ and $\{\widehat{\rho}_{j+1}\}_{j\in\mathbb{Z}^+}$ are all Gaussian distributions.*

As a consequence these distributions can be entirely characterized by their mean and covariance. We write

$$\begin{aligned}\widehat{\rho}_{j+1} &= \mathbb{P}(v_{j+1}|Y_j) = N(\widehat{m}_{j+1}, \widehat{C}_{j+1}) && \text{(prediction)}\\ \rho_{j+1} &= \mathbb{P}(v_{j+1}|Y_{j+1}) = N(m_{j+1}, C_{j+1}) && \text{(analysis)}\end{aligned}$$

and aim to find update formulae for these means and covariances. The Kalman filter algorithm achieves this.

Theorem 9.3 (Kalman filter)**.** *Under Assumption 9.1 and if* $\Sigma, \Gamma, C_0 > 0$ *(positive definite) it follows that, for all* $j \in Z^+$*,* $C_j > 0$ *and*

$$\widehat{m}_{j+1} = Mm_j$$

$$\widehat{C}_{j+1} = MC_jM' + \Sigma$$

$$C_{j+1}^{-1} = (MC_jM^T + \Sigma)^{-1} + H^T\Gamma^{-1}H$$

$$C_{j+1}^{-1}m_{j+1} = (MC_jM^T + \Sigma)^{-1}Mm_j + H^T\Gamma^{-1}y_{j+1}$$

Proof. The proof proceeds by breaking the Kalman filter step above into the prediction and the analysis steps.
Prediction: The mean and variance of the prediction step may be calculated as follows. The mean is given by:

$$\begin{aligned}\widehat{m}_{j+1} &= \mathbb{E}\left[v_{j+1}|Y_j\right]\\ &= \mathbb{E}\left[Mv_j + \xi_j|Y_j\right] && \text{since } v_{j+1} = Mv_j + \xi_j\\ &= M\,\mathbb{E}\left[v_j|Y_j\right] + \mathbb{E}\left[\xi_j|Y_j\right]\\ &= Mm_j && \text{since } \xi_j \text{ and } Y_j \text{ are independent.}\end{aligned}$$

The covariance is given by:

$$\begin{aligned}\widehat{C}_{j+1} =& \mathbb{E}\left[(v_{j+1} - \widehat{m}_{j+1}) \otimes (v_{j+1} - \widehat{m}_{j+1})|Y_j\right]\\ =& \mathbb{E}\left[M(v_j - m_j) \otimes M(v_j - m_j)|Y_j\right] + \mathbb{E}\left[\xi_j \otimes \xi_j|Y_j\right]\\ &+ \mathbb{E}\left[\xi_j \otimes M(v_j - m_j)|Y_j\right] + \mathbb{E}\left[M(v_j - m_j) \otimes \xi_j|Y_j\right]\\ =& M\,\mathbb{E}\left[(v_j - m_j) \otimes (v_j - m_j)|Y_j\right]M^T + \Sigma && \text{since } \xi_j \text{ and } v_j \text{ are independent}\\ =& MC_jM^T + \Sigma.\end{aligned}$$

Thus in the linear Gaussian setting the prediction operator P from $\rho_j = N(m_j, C_j)$ to $\widehat{\rho}_{j+1} = N(\widehat{m}_{j+1}, \widehat{C}_{j+1})$ is given by

$$\begin{aligned}\widehat{m}_{j+1} &= Mm_j,\\ \widehat{C}_{j+1} &= MC_jM^T + \Sigma.\end{aligned}$$

Analysis: The analysis step may be derived as follows, using Bayes' Theorem:

$$\begin{aligned}\mathbb{P}(v_{j+1}|Y_{j+1}) &= \mathbb{P}(v_{j+1}|y_{j+1}, Y_j) \\ &\propto \mathbb{P}(y_{j+1}|v_{j+1}, Y_j)\mathbb{P}(v_{j+1}|Y_j) \\ &= \mathbb{P}(y_{j+1}|v_{j+1})\mathbb{P}(v_{j+1}|Y_j)\end{aligned}$$

This gives us:

$$\begin{aligned}\exp\Big(-\frac{1}{2}|v_{j+1} - m_{j+1}|^2_{C_{j+1}}\Big) &\propto \exp\Big(-\frac{1}{2}|y_{j+1} - Hv_{j+1}|^2_\Gamma\Big)\exp\Big(-\frac{1}{2}|v_{j+1} - \widehat{m}_{j+1}|^2_{\widehat{C}_{j+1}}\Big) \\ &= \exp\Big(-\frac{1}{2}|y_{j+1} - Hv_{j+1}|^2_\Gamma - \frac{1}{2}|v_{j+1} - \widehat{m}_{j+1}|^2_{\widehat{C}_{j+1}}\Big)\end{aligned}$$

Taking logs and matching quadratic and linear terms in v_{j+1} from either side of this identity gives the update operator L_j from $\widehat{\rho}_{j+1} = N(\widehat{m}_{j+1}, \widehat{C}_{j+1})$ to $\rho_{j+1} = N(m_{j+1}, C_{j+1})$:

$$\begin{aligned}C^{-1}_{j+1} &= \widehat{C}^{-1}_{j+1} + H^T\Gamma^{-1}H \\ C^{-1}_{j+1}m_{j+1} &= \widehat{C}^{-1}_{j+1}\widehat{m}_{j+1} + H^T\Gamma^{-1}y_{j+1}\end{aligned}$$

Combining the prediction operator P and update operator L_j gives us the desired update formulae.

Positive-definiteness of C_j: It remains to show that $C_j > 0$ for all $j \in Z^+$. We will use induction. By assumption the result holds true for $j = 0$. Assume that it is true for C_j. From the prediction operator P we have, for $u \neq 0$,

$$\begin{aligned}\langle u, \widehat{C}_{j+1}u\rangle &= \langle u, MC_jM^Tu\rangle + \langle u, \Sigma u\rangle \\ &= \langle C_j(M^Tu), (M^Tu)\rangle + \langle u, \Sigma u\rangle \\ &\geq \langle u, \Sigma u\rangle \quad \text{since } C_j > 0 \\ &> 0 \quad \text{since } \Sigma > 0.\end{aligned}$$

So $\widehat{C}_{j+1}, \widehat{C}^{-1}_{j+1} > 0$. Then from update operator L_j:

$$\begin{aligned}\langle u, C^{-1}_{j+1}u\rangle &= \langle u, \widehat{C}^{-1}_{j+1}u\rangle + \langle u, H^T\Gamma^{-1}Hu\rangle \\ &= \langle u, \widehat{C}^{-1}_{j+1}u\rangle + \langle \Gamma^{-1}(Hu), (Hu)\rangle \\ &\geq \langle u, \widehat{C}^{-1}_{j+1}u\rangle \quad \text{since } \Gamma > 0 \\ &> 0 \quad \text{since } \widehat{C}^{-1}_{j+1} > 0.\end{aligned}$$

Therefore, $C_{j+1}, C^{-1}_{j+1} > 0$, which concludes the proof. □

Remark 9.4. The previous proof reveals two interesting facts about the structure of the Kalman filter updates. The first is that the covariance update is independent of the observed data. The second is that the update formulae for the covariance are linear in the prediction step, but nonlinear in the analysis step; specifically the analysis step is linear in the precisions (inverse covariances). □

9.3 Kalman Filter: Alternative Formulation

We now rewrite the Kalman filter in an alternative form. This formulation is written in terms of the covariances directly, and does not involve the precisions. Furthermore, the formulation in the previous section requires a matrix inversion in the state space while this one requires only inversion the data space, namely S_{j+1}^{-1} in what follows. In many applications the observation space dimension is much smaller than the state space dimension ($m \ll n$), and then the formulation given in this section is much cheaper to compute than that which arises from the derivation in the previous section.

Corollary 9.5 (Standard Form of Kalman Filter)**.** *Under the same assumptions as Theorem 9.3, the Kalman update formulae may be written as*

$$m_{j+1} = \widehat{m}_{j+1} + K_{j+1} d_{j+1}$$
$$C_{j+1} = (I - K_{j+1} H)\widehat{C}_{j+1}$$

where

$$d_{j+1} = y_{j+1} - H\widehat{m}_{j+1}$$
$$S_{j+1} = H\widehat{C}_{j+1} H^T + \Gamma$$
$$K_{j+1} = \widehat{C}_{j+1} H^T S_{j+1}^{-1}$$

and where $\widehat{m}_{j+1}, \widehat{C}_{j+1}$ are defined as in the proof of Theorem 9.3.

Remark 9.6. The vector d_{j+1} is known as the *innovation* and the matrix K_{j+1} as the *Kalman gain*. Note that d_{j+1} measures the mismatch of the predicted state from the given data.

The corollary may be proved by an application of the following:

Lemma 9.7. *(Woodbury Matrix Identity) Let $A \in R^{p\times p}$, $U \in R^{p\times q}$, $C \in R^{q\times q}$, $V \in R^{q\times p}$. If $A, C > 0$, then $A + UCV$ is invertible and*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

Combining the form of d_{j+1} and $\widehat{m}_{j+1}$ shows that the update formula for the Kalman mean can be written as follows:

$$m_{j+1} = (I - K_{j+1}H)\widehat{m}_{j+1} + K_{j+1}y_{j+1}, \quad \widehat{m}_{j+1} = Mm_j. \tag{9.1}$$

This update formula has the very natural interpretation that the mean update is formed as a linear combination of the evolution of the noise-free dynamics and of the data. It may be derived from an optimization perspective, the topic of the next subsection.

9.4 Optimization Perspective: Mean of Kalman Filter

Since ρ_{j+1} is Gaussian,

$$m_{j+1} = \text{argmax}_m \rho_{j+1}(m)$$

Using the update operator L_j defined by the first two displays in Theorem 9.3, we have:

$$m_{j+1} = \text{argmax}_m \exp\left(-\frac{1}{2}|y_{j+1} - Hm|^2_\Gamma - \frac{1}{2}|m - \widehat{m}_{j+1}|^2_{\widehat{C}_{j+1}}\right)$$

which can be rewritten as:

$$m_{j+1} = \text{argmin}_m \left(\frac{1}{2}|y_{j+1} - Hm|^2_\Gamma + \frac{1}{2}|m - \widehat{m}_{j+1}|^2_{\widehat{C}_{j+1}}\right)$$

In other words, m_{j+1} is chosen to fit both the observed data y_{j+1} and the predictions $\widehat{m}_{j+1}$ as well as possible. The covariances Γ and $\widehat{C}_{j+1}$ act to determine the relative weighting between the two quadratic terms. The solution of the minimization problem is given by (9.1).

To see this we define

$$\mathsf{I}(m) := \frac{1}{2}|y_{j+1} - Hm|^2_\Gamma + \frac{1}{2}|m - \widehat{m}_{j+1}|^2_{\widehat{C}_{j+1}}$$

and write $v' = m - \widehat{m}_{j+1}$, $y' = y_{j+1} - H\widehat{m}_{j+1}$ and $C' = \widehat{C}_{j+1}$. Then the minimization problem may be reformulated as

$$m_{j+1} = \widehat{m}_{j+1} + \text{argmin}_{v'} \left(\frac{1}{2}|y' - Hv'|^2_\Gamma + \frac{1}{2}\langle v', b\rangle\right)$$

where the minimization is subject to the constraint $C'b = v'$. Using Lagrange multipliers we write

$$\mathsf{J}(v') = \frac{1}{2}|y' - Hv'|^2_\Gamma + \frac{1}{2}\langle v', b\rangle + \langle \lambda, C'b - v'\rangle; \tag{9.2}$$

computing the dervative and setting to zero gives

$$\begin{aligned} -H^T\Gamma^{-1}(y' - Hv') + \frac{1}{2}b - \lambda &= 0, \\ \frac{1}{2}v' + C'\lambda &= 0, \\ v' - C'b &= 0. \end{aligned}$$

The last two equations imply that $C'(2\lambda + b) = 0$. Thus we set $\lambda = -\frac{1}{2}b$ and drop the second equation, replacing the first by

$$-H^T\Gamma^{-1}(y' - HC'b) + b = 0.$$

Solving this for b gives

$$\begin{aligned} m &= \widehat{m}_{j+1} + v' \\ &= \widehat{m}_{j+1} + C'b \\ &= \widehat{m}_{j+1} + C'(H^T\Gamma^{-1}HC' + I)^{-1}H^T\Gamma^{-1}y' \\ &= \widehat{m}_{j+1} + C'(H^T\Gamma^{-1}HC' + I)^{-1}H^T\Gamma^{-1}(y_{j+1} - H\widehat{m}_{j+1}) \\ &= (I - K_{j+1}H)\widehat{m}_{j+1} + K_{j+1}y_{j+1} \end{aligned}$$

where we have defined

$$K_{j+1} = C'(H^T\Gamma^{-1}HC' + I)^{-1}H^T\Gamma^{-1}.$$

It remains to show that K_{j+1} agrees with the prescription given in Corollary 9.5. To see this we note that if we choose S to be any matrix satisfying $K_{j+1} = C'H^TS^{-1}$ then

$$H^TS^{-1} = (H^T\Gamma^{-1}HC' + I)^{-1}H^T\Gamma^{-1}$$

so that

$$(H^T\Gamma^{-1}HC' + I)H^T = H^T\Gamma^{-1}S.$$

Thus

$$H^T\Gamma^{-1}HC'H^T + H^T = H^T\Gamma^{-1}S$$

which may be achieved by choosing any S so that

$$\Gamma^{-1}(HC'H^T + \Gamma) = \Gamma^{-1}S$$

and multiplication by Γ gives the desired formula for S.

9.5 Optimality of Kalman Filter

The following theorem states that the Kalman filter gives the best estimator of the mean in an online setting. In the following $\mathbb{E}$ denotes expectation with respect to all randomness present in the problem statement, through the initial condition, the noisy dynamical evolution, and the noisy data. Furthermore $\mathbb{E}[\cdot|Y_j]$ denotes conditional expectation, given the data Y_j upto time j.

Theorem 9.8 (Optimality of Kalman Filter)**.** *Let $\{m_j\}$ be the sequence computed using the Kalman filter, and $\{z_j\}$ be any sequence in $\mathbb{R}^n$ such that z_j is Y_j measurable.*[3] *Then:*

$$\forall j \in N, \quad \mathbb{E}\left[|v_j - m_j|^2 \mid Y_j\right] \leq \mathbb{E}\left[|v_j - z_j|^2 \mid Y_j\right]$$

.

[3]For practical purposes this means z_j is a fixed non-random function of given observed Y_j.

Proof. Note that m_j and z_j are fixed and non-random, given Y_j. Thus we have:

$$\begin{aligned}
\mathbb{E}\left[|v_j - z_j|^2 \mid Y_j\right] &= \mathbb{E}\left[|v_j - m_j + m_j - z_j|^2 \mid Y_j\right] \\
&= \mathbb{E}\left[|v_j - m_j|^2 + 2\Big\langle v_j - m_j, m_j - z_j\Big\rangle + |m_j - z_j|^2 \mid Y_j\right] \\
&= \mathbb{E}\left[|v_j - m_j|^2 \mid Y_j\right] + 2\Big\langle \mathbb{E}\left[v_j - m_j \mid Y_j\right], m_j - z_j\Big\rangle + |m_j - z_j|^2 \\
&= \mathbb{E}\left[|v_j - m_j|^2 \mid Y_j\right] + 2\Big\langle \mathbb{E}\left[v_j \mid Y_j\right] - m_j, m_j - z_j\Big\rangle + |m_j - z_j|^2 \\
&= \mathbb{E}\left[|v_j - m_j|^2 \mid Y_j\right] + 0 + |m_j - z_j|^2 \\
&\geq \mathbb{E}\left[|v_j - m_j|^2 \mid Y_j\right]
\end{aligned}$$

The fifth step follows since $m_j = \mathbb{E}\left[v_j \mid Y_j\right]$. □

9.6 Discussion and Bibliography

The original paper of Kalman, which is arguably the first systematic presentation of a methodology to combine models with data, is [57]. The continuous time analogue of that work may be found in [58]. The book [45] overviews the subject in the contextof time-series analysis and economics. The proof of Corollary 9.5 may be found in [70]. The paper [96] contains an application of the optimality property of the Kalman filter (which applies beyond the linear Gaussian setting to the mean of the filtering distribution in quite general settings.)

10 Optimization for FItering and Smoothing: 3DVAR and 4DVAR

In the previous chapter we showed how the mean of the Kalman filter could be derived through an optimization principle, once the predictive covariance is known. In this chapter we discuss two optimization based approaches to filtering and smoothing, namely the 3DVAR and 4DVAR methodologies. We emphasize that the methods we present in this chapter do not provide approximations of the filtering and smoothing probability distributions; they simply provide estimates of the signal, given data, in the filtering (on-line) and smoothing (off-line) data scenarios.

10.1 The Setting

3DVAR borrows from the Kalman filter optimization principle, but substitutes a fixed given covariance for the predictive covariance. Here "VAR" refers to variational, and encodes the concept of optimization. The 3D and 4D, respectively, refer to three Euclidean spatial dimensions and to three Euclidean spatial dimensions plus a time dimension; this nomenclature reflects the historical derivation of these problems in the geophysical sciences, but the specific structure of fields over three dimensional Euclidean space plays no role in the generalized form of the methods described here. The key distinction is that 3DVAR solves a sequence of optimization problems at each point in time (hence is an on-line filtering method); in contrast 4DVAR solves an optimization problem which involves data distributed over time (and is an off-line smoothing method).

Throughout we consider the set-up in which the dynamics model is nonlinear, but the observation operator is linear, commonly occurring in applications. We thus have a discrete-time dynamical system with noisy state transitions and noisy observations given by

$$\begin{aligned}
\text{Dynamics Model:} \quad & v_{j+1} = \Psi(v_j) + \xi_j, \quad j \in \mathbb{Z}^+ \\
\text{Data Model:} \quad & y_{j+1} = Hv_{j+1} + \eta_{j+1}, \quad j \in \mathbb{Z}^+ \\
\text{Probabilistic Structure:} \quad & v_0 \sim N(m_0, C_0), \quad \xi_j \sim N(0, \Sigma), \quad \eta_j \sim N(0, \Gamma) \\
\text{Probabilistic Structure:} \quad & v_0 \perp \{\xi_j\} \perp \{\eta_j\} \text{ independent}
\end{aligned}$$

Throughout this chapter $H \in \mathbb{R}^{m \times n}$.

10.2 3DVAR

We introduce 3DVAR by analogy with the update formula (9.1) for the Kalman filter, and its derivation through optimization from section 9.4. The primary differences between 3DVAR and the Kalman filter mean update is that $\Psi(\cdot)$ can be nonlinear for 3DVAR, and that for 3DVAR we have no closed update formula for the covariances. To deal with this second issue 3DVAR uses a fixed predicted covariance, independent of time j, and pre-specified. The resulting minimization problem, and its solution, is described in Table 10.1, making the analogy with the Kalman filter. Note that the minimization itself is of a quadratic functional, and so may be solved by means of linear algebra. The constraint formulation used for the Kalman filter, in section 9.4, may also be applied and used to derive the mean update formula.

Kalman Filter	3DVAR
$m_{j+1} = \arg\min_m \mathsf{J}(m)$	$m_{j+1} = \arg\min_m \mathsf{J}(m)$
$\mathsf{J}(m) = \frac{1}{2}\lvert y_{j+1} - Hm\rvert^2_\Gamma + \frac{1}{2}\lvert m - \widehat{m}_{j+1}\rvert^2_{\widehat{C}_{j+1}}$	$\mathsf{J}(m) = \frac{1}{2}\lvert y_{j+1} - Hm\rvert^2_\Gamma + \frac{1}{2}\lvert m - \widehat{m}_{j+1}\rvert^2_{\widehat{C}}$
$\widehat{m}_{j+1} = Mm_j$	$\widehat{m}_{j+1} = \Psi(m_j)$
$m_{j+1} = (I - K_{j+1}H)\widehat{m}_{j+1} + K_{j+1}y_{j+1}$	$m_{j+1} = (I - KH)\widehat{m}_{j+1} + Ky_{j+1}$

Table 10.1 Comparison of Kalman Filter and 3DVAR update formulae

The Kalman gain K for 3DVAR is fixed, because the predicted covariance $\widehat{C}$ is fixed. By analogy with Corollary 9.5 we have the following formulae for the 3DVAR gain matrix K, and the update formula for the estimator m_j:

$$\begin{aligned} S &= H\widehat{C}H^T + \Gamma \\ K &= \widehat{C}H^T S^{-1} \\ m_{j+1} &= (I - KH)\Psi(m_j) + Ky_{j+1}. \end{aligned}$$

The method also delivers an implied analysis covariance $C = (I - KH)\widehat{C}$. Note that the resulting algorithm which maps m_j to m_{j+1} may be specified directly in terms of the gain K, without need to introduce $\widehat{C}, C$ and S. In the remainder of this section we simply view K as fixed and given. In this setting we show that the 3DVAR algorithm produces accurate state estimation under vanishing noise assumptions in the dynamics/data model. The governing assumptions concerning the dynamics/data model are encapsulated in:

Assumption 10.1. *Consider the dynamics/data model under the assumptions that* $\xi_j \equiv 0, \Gamma = \gamma^2\Gamma_0, |\Gamma_0| = 1$ *and assume that the data* y_{j+1} *used in the 3DVAR algorithm is found from observing a true signal* $v_j^\dagger$ *given by*

$$\begin{aligned} \textit{Dynamics Model:} \quad & v_{j+1}^\dagger = \Psi(v_j^\dagger), \quad j \in \mathbb{Z}^+ \\ \textit{Data Model:} \quad & y_{j+1} = Hv_{j+1}^\dagger + \gamma\eta_{j+1,0}^\dagger, \quad j \in \mathbb{Z}^+. \end{aligned}$$

With this assumption of noise-free dynamics ($\xi_j \equiv 0$) we deduce that the 3DVAR filter produces output which, asymptotically, has an error of the same size as the observational noise error γ. The key additional assumption in the theorem that allows this deduction is a relationship between the Kalman gain K and the derivative $D\Psi(\cdot)$ of the dynamics model. Encoded in the assumption are two ingredients: that the observation operator H is rich enough in principle to learn enough components of the system to synchronize the whole system; and that K is designed cleverly enough to effect this synchronization. The proof of the theorem is simply using these two ingredients and then controlling the small stochastic perturbations, arising from Assumption 10.1.

Theorem 10.2 (Accuracy of 3DVAR). *Let Assumption 10.1 hold with $\eta^\dagger_{j,0} \sim N(0, \Gamma_0)$. an i.i.d. sequence. Assume that, for the gain matrix K appearing in the 3DVAR method, there exists a norm $\|\cdot\|$ on $\mathbb{R}^n$ and constant $\lambda \in (0,1)$ s.t. $\|(I-KH)D\Psi(v)\| \leq \lambda \; \forall v \in \mathbb{R}^n$. Then there is constant $c \in (0,\infty)$ such that the 3DVAR algorithm satisfies the following large-time asymptotic error bound:*

$$\limsup_{j\to\infty} \mathbb{E}\|m_j - v_j^\dagger\| \leqslant \frac{c\gamma}{1-\lambda},$$

where the expectation is taken with respect to the sequence $\{\eta^\dagger_{j,0}\}$.

Proof. We have

$$\begin{aligned} v^\dagger_{j+1} &= \Psi(v^\dagger_j), \\ m_{j+1} &= (I-KH)\Psi(m_j) + Ky_{j+1} \end{aligned}$$

and hence that

$$\begin{aligned} v^\dagger_{j+1} &= (I-KH)\Psi(v^\dagger_j) + KH\Psi(v^\dagger_j), \\ m_{j+1} &= (I-KH)\Psi(m_j) + KH\Psi(v^\dagger_j) + \gamma K\eta^\dagger_{j+1,0}. \end{aligned}$$

Define $e_j = m_j - v^\dagger_j$. By subtracting the evolution equation for $v^\dagger_j$ from that for m_j we obtain, using the mean value theorem,

$$\begin{aligned} e_{j+1} &= m_{j+1} - v^\dagger_{j+1} \\ &= (I-KH)(\Psi(m_j) - \Psi(v^\dagger_j)) + \gamma K\eta^\dagger_{j+1,0} \\ &= \left((I-KH)\int_0^1 D\Psi\Big(sm_j + (1-s)v^\dagger_j\Big)ds\right)e_j + \gamma K\eta^\dagger_{j+1,0}. \end{aligned}$$

As a result, by the triangle inequality,

$$\begin{aligned} \|e_{j+1}\| &\leq \left\|\left(\int_0^1 (I-KH)D\Psi\Big(sm_j + (1-s)v^\dagger_j\Big)ds\right)e_j\right\| + \|\gamma K\eta^\dagger_{j+1,0}\| \\ &\leq \left(\int_0^1 \Big\|(I-KH)D\Psi\Big(sm_j + (1-s)v^\dagger_j\Big)\Big\|ds\right)\|e_j\| + \|\gamma K\eta^\dagger_{j+1,0}\| \\ &\leq \lambda\|e_j\| + \gamma\|K\eta^\dagger_{j+1,0}\|. \end{aligned}$$

Taking expectations on both sides, we obtain, for $c := \mathbb{E}\|K\eta^\dagger_{0,j+1}\| > 0$,

$$\begin{aligned} \mathbb{E}\|e_{j+1}\| &\leq \lambda\,\mathbb{E}\|e_j\| + \gamma\,\mathbb{E}\|K\eta^\dagger_{j+1,0}\| \\ &\leq \lambda\,\mathbb{E}\|e_j\| + \gamma c. \end{aligned} \tag{10.1}$$

Using the Discrete Time Gronwall Lemma, we have that:

$$\begin{aligned} \mathbb{E}\|e_j\| &\leq \lambda^j\,\mathbb{E}\|e_0\| + \sum_{\ell=0}^{j-1} c\lambda^\ell\gamma \\ &\leq \lambda^j\,\mathbb{E}\|e_0\| + c\gamma\frac{1-\lambda^j}{1-\lambda}. \end{aligned} \tag{10.2}$$

where $e_0 = m_0 - v_0$. As $\lambda < 1$, the desired statement follows. □

10.3 4DVAR

Recall that 3DVAR differs from 4DVAR because, whilst also based on an optimization principle, 4DVAR is applied in a distributed fashion over all data in the time interval $j = 1, \cdots, J$; in contrast 3DVAR is applied sequentially from time $j-1$ to time j, for $j = 1, \cdots, J$. We consider two forms of the methodology: *weak constraint 4DVAR (w4DVAR)*, in which the fact that the dynamics model contains randomness is allowed for in the optimization; and *4DVAR* (sometimes known as *strong constraint 4DVAR*) which can be derived from w4DVAR in the limit of $\Sigma \to 0$ (no randomness in the dynamics).

The objective function minimized in w4DVAR is

$$\mathsf{J}(v) = \frac{1}{2}|v_0 - m_0|^2_{C_0} + \sum_{j=0}^{J-1} \frac{1}{2}|v_{j+1} - \Psi(v_j)|^2_{\Sigma} + \sum_{j=0}^{J-1} \frac{1}{2}|y_{j+1} - Hv_j|^2_{\Gamma}$$

where $v = \{v_j\}_{j=0}^{J} \in \mathbb{R}^{(J+1)n}$, $y = \{y_j\}_{j=1}^{J} \in \mathbb{R}^{Jn}$, $v_j \in \mathbb{R}^n$, $y_j \in \mathbb{R}^m$, H is the observation operator, Σ is the random dynamical system covariance, Γ is the data noise covariance and m_0 and C_0 are the mean and covariance of the initial state. The three terms in the objective function enforce, in turn, the initial condition v_0, the dynamics model and the data model. Note that, because Ψ is nonlinear, the optimization is not-quadratic and cannot be solved in closed form. In contrast, each step of 3DVAR required solution of a quadratic optimization, tractable in closed form.

Theorem 10.3 (Minimizer Exists for w4DVAR)**.** *Under the assumption that Ψ is bounded and continuous with $|\Psi(u)|_{\Sigma} \leq M$ J has a minimizer v^* which is a MAP estimator for the smoothing problem.*

Proof. Note that, for all $v \in \mathbb{R}^{(J+1)n}$,

$$\mathsf{J}(v) \geq 0,$$
$$\mathsf{J}(0) \leq \frac{1}{2}|m_0|^2_{C_0} + \frac{J}{2}M^2 + \frac{1}{2}|y|^2.$$

Thus an infimum $\bar{\mathsf{J}}$ exists, and so it remains to show that it is obtained at some $v^* \in \mathbb{R}^{(J+1)n}$. Let $v^{(k)}$ be an infimizing sequence. Then for any $\delta > 0$ there is $K = K(\delta)$ such that, for all $k \geq K$,

$$\frac{1}{2}|v_0^{(k)} - m_0|^2_{C_0} + \sum_{j=0}^{J-1} \frac{1}{2}|v_{j+1}^{(k)} - \Psi(v_j^{(k)})|^2_{\Sigma} + \sum_{j=0}^{J-1} \frac{1}{2}|y_{j+1} - Hv_j^{(k)}|^2_{\Gamma} = \mathsf{J}(v^{(k)}) \leq \bar{\mathsf{J}} + \delta.$$

From this we deduce that

$$|v_0^{(k)}|^2_{C_0} + \sum_{j=0}^{J-1} |v_{j+1}^{(k)}|^2_{\Sigma} \leq \bar{\mathsf{J}} + \delta + |m_0|^2_{C_0} + JM^2.$$

This implies that $|v^{(k)}| \leq C$, for some constant C independent of k. Therefore there is a convergent subsequence with limit v^*. Since

$$\bar{\mathsf{J}} \leq \mathsf{J}(v^{(k)}) \leq \bar{\mathsf{J}} + \delta$$

we deduce by continuity that

$$\bar{\mathsf{J}} \leq \frac{1}{2}|v_0^* - m_0|^2_{C_0} + \sum_{j=0}^{J-1} \frac{1}{2}|v_{j+1}^* - \Psi(v_j^*)|^2_\Sigma + \sum_{j=0}^{J-1} \frac{1}{2}|y_{j+1} - Hv_j^*|^2_\Gamma \leq \bar{\mathsf{J}} + \delta$$

for arbitrary $\delta > 0$. It follows that the infimum is attained at v^*.

Finally we note that

$$\mathsf{J}(v) = l(v;\eta) + r(v)$$

with the notation as in the smoothing problem from section 8.2, specialized to the case $h(\cdot) = H\cdot$. Hence the infinimizer is also a MAP estimator. □

We now consider the vanishing dynamical noise limit of w4DVAR. This is to minimize

$$\mathsf{J}_0(v) = \frac{1}{2}|v_0 - m_0|^2_{C_0} + \sum_{j=0}^{J-1} \frac{1}{2}|y_{j+1} - Hv_j|^2_\Gamma$$

subject to the hard constraint that

$$v_{j+1} = \Psi(v_j), \quad j = 0, \cdots, J-1.$$

This is 4DVAR. Note that by using the constraint 4DVAR can be written as a minimization over v_0, rather than over the entire sequence $\{v_j\}_{j=0}^J$ as is required in w4DVAR.

We let J_σ denote the objective function J from w4DVAR in the case where $\Sigma \to \sigma^2\Sigma_0$. Roughly speaking, the following result shows that minimizers of J_σ converge as $\sigma \to 0$ to points in $\mathbb{R}^{(J+1)n}$ which satisfy the hard constraint associated with 4DVAR.

Theorem 10.4 (Small Signal Noise Limit of w4DVAR)**.** *Let $v^{(\sigma)}$ be the minimizer of J_σ. Under the assumption that Ψ is bounded and continuous with $|\Psi(u)|_\Sigma \leq M$ then as $\sigma \to 0$ there is a convergent subsequence of $v^{(\sigma)}$ with limit v^* satisfying $v_{j+1}^* = \Psi(v_j^*)$.*

Proof. Throughout this proof C is a constant which may change from instance to instance, but is independent of σ. Consider $v \in \mathbb{R}^{(J+1)n}$ defined by $v_0 = m_0$ and $v_{j+1} = \Psi(v_j)$. Then v is bounded, as $\Psi(\cdot)$ is bounded, and the bound is independent of σ. Furthermore

$$\mathsf{J}_\sigma(v) = \sum_{j=0}^{J-1} \frac{1}{2}|y_{j+1} - Hv_{j+1}|^2_\Gamma \leq C.$$

where C is independent of σ. It follows that

$$\mathsf{J}_\sigma(v^{(\sigma)}) \leq \mathsf{J}_\sigma(v) \leq C.$$

Thus

$$\begin{aligned} &\frac{1}{2}|v_{j+1}^{(\sigma)} - \Psi(v_j^{(\sigma)})|^2_\Sigma \leq C \\ \implies &\frac{1}{2}|v_{j+1}^{(\sigma)} - \Psi(v_j^{(\sigma)})|^2_{\Sigma_0} \leq \sigma^2 C \\ \text{and} \quad &\frac{1}{2}|v_0^{(\sigma)} - m_0|^2_{C_0} \leq C \end{aligned}$$

Since Ψ is bounded these bounds imply that $|v^{(\sigma)}| \leq C$. Therefore there is a limit $v^* : v^{(\sigma)} \to v^*$ along a subsequence. By continuity

$$0 \leq \frac{1}{2}|v^*_{j+1} - \Psi(v^*_j)|^2_{\Sigma_0} \leftarrow \frac{1}{2}|v^{(\sigma)}_{j+1} - \Psi(v^{(\sigma)}_j)|^2_{\Sigma_0} \leq \sigma^2 C$$

Therefore, letting $\sigma \to 0$:

$$\frac{1}{2}|v^*_{j+1} - \Psi(v^*_j)|^2_{\Sigma_0} = 0 \implies v^*_{j+1} = \Psi(v^*_j)$$

□

10.4 Discussion and Bibliography

The 3DVAR and 4DVAR methodologies, in the context of weather forecasting, are discussed in [72] and [37] respectively. The accuracy analysis presented here is similar to that which first appeared in the papers [14], [83] and was developed further in [69, 68]. It arises from considering stochastic perturbations of the seminal work of Titi and co-workers, exemplified by the paper [47]. For an overview of variational methods, and their links to problems in physics and mechanics, see the book [1], and the references therein; see also the paper [15].

11 Particle Filter

This chapter is devoted to the particle filter, a method designed to approximate the true filtering distribution by a sum of Dirac measures, in a provably convergent fashion, as the number of Dirac measures approaches infinity. In particular we introduce the bootstrap filter, a method also known as sequential importance resampling; it is linked to the material on importance sampling described in chapter 6. We note that the Kalman filter completely characterizes the filtering distribution, but only in the linear Gaussian setting; the methodology that we introduce here is very general and applies in nonlinear, non-Gaussian settings.

11.1 Introduction

In this chapter we return to the setting in which we introduced filtering and smoothing, with nonlinear stochastic dynamical system and nonlinear observation operator, namely the model

$$\begin{aligned} v_{j+1} &= \Psi(v_j) + \xi_j & \xi_j &\sim N(0,\Sigma) \text{ i.i.d.}, \\ y_{j+1} &= h(v_{j+1}) + \eta_j & \eta_j &\sim N(0,\Gamma) \text{ i.i.d.}, \end{aligned}$$

with $v_0 \sim N(m_0, C_0)$ independent of the independent i.i.d. sequences $\{\xi_j\}$ and $\{\eta_j\}$. Here $\Psi(\cdot)$ drives the dynamical system and $h(\cdot)$ is the observation operator. Recall that we denote by $Y_j = \{y_1, \ldots, y_j\}$ all the data up to time j and by ρ_j the pdf of $v_j|Y_j$ that is $\rho_j = \mathbb{P}(v_j|Y_j)$. Assuming $\rho_0 = N(m_0, C_0)$, the filtering problem is to determine ρ_{j+1} from ρ_j. We may do so in two steps: first, we run forward the Markov chain generated by the stochastic dynamical system (prediction), and second, we incorporate the data by an application of Bayes Theorem (analysis).

For the prediction step, we define the operator P acting on a pdf ρ as an application of a Markov kernel defined by

$$(P\rho)(v) = \int_{\mathbb{R}^n} p(u,v)\rho(u)du \tag{11.1}$$

where $p(u,v)$ is the associated pdf of the stochastic dynamics, so that

$$p(u,v) = \frac{1}{\sqrt{(2\pi)^n \det\Sigma}} \exp\left(-\frac{1}{2}|v - \Psi(u)|^2_\Sigma\right).$$

Thus we obtain

$$\mathbb{P}(v_{j+1}|Y_j) = \widehat{\rho}_{j+1} = P\rho_j.$$

We then define the analysis operator L_j acting on a pdf ρ to correspond to an application of Bayes Theorem, namely

$$(L_j\rho)(v) = \frac{\exp(-\frac{1}{2}|y_{j+1} - h(v)|^2_\Gamma)\rho(v)}{\int_{\mathbb{R}^m} \exp(-\frac{1}{2}|y_{j+1} - h(v)|^2_\Gamma)\rho(v)dv} \tag{11.2}$$

and so finally we obtain

$$\rho_{j+1} = L_j\widehat{\rho}_{j+1} = L_jP\rho_j.$$

We now describe a way to numerically approximate, and update, the pdfs ρ_j.

11.2 The Bootstrap Particle Filter

The bootstrap particle filter (BPF) can be though of as performing sequential importance resampling. Let S^N be an operator acting on a pdf ρ by producing an $N-$samples Dirac approximation of ρ, that is

$$(S^N\rho)(u) = \sum_{n=1}^{N} w_j \delta(u - v^{(n)})$$

where $v^{(1)}, \ldots, v^{(N)}$ are i.i.d samples from ρ that are weighted uniformly i.e. $w_j = \frac{1}{N}$. Note that $S^N\rho = \rho_{MC}^N$ given by equation (6.3). We will use the operator S^N to approximate the measure produced by the Markov kernel step P within the overall filtering map L_jP. Note that S_N is a *random* map taking pdfs into pdfs if we interpret weighted sums of Diracs as a special pdf.

Let $\rho_0^N = \rho_0 = N(m_0, C_0)$ and let ρ_j^N denote a particle approximation of the pdf ρ_j that we will determine in what follows. We define

$$\widehat{\rho}_{j+1}^N = S^N P \rho_j^N;$$

this is an approximation of $\widehat{\rho}_{j+1}$ from the previous section. We then apply the operator L_j to act on $\widehat{\rho}_{j+1}^N$ by appropriately reconfiguring the weights w_j according to the data.

To understand this reconfiguration of the weights we use the fact that, if

$$(L\rho)(v) = \frac{g(v)\rho(v)}{\int_{\mathbb{R}^m} g(v)\rho(v)dv}$$

and if

$$\rho(v) = \frac{1}{N}\sum_{n=1}^{N} \delta(v - v^{(n)})$$

then

$$(L\rho)(v) = \sum_{n=1}^{N} w^{(n)} \delta(v - v^{(n)})$$

where

$$\bar{w}^{(n)} = g(v^{(n)})$$

and the $w^{(n)}$ are found from the $\bar{w}^{(n)}$ by renormalizing them to sum to one. We use this calculation concerning the application of Bayes formula to sums of Diracs within the following desired approximation of the filtering update formula:

$$\rho_{j+1} \approx \rho_{j+1}^N = L_j \widehat{\rho}_{j+1}^N = L_j S^N P \rho_j^N.$$

The steps for the method are summarized in Algorithm 11.1.

Algorithm 11.1 Bootstrap Particle Filter

1: **Input**: Initial distribution $\rho_0^N = \rho_0$, observations Y_J, number of particles N
2: **Particle Generation**: For $j = 0, 1, \ldots, J-1$, perform

1. Draw $v_j^{(n)} \sim \rho_j^N$ for $n = 1, \ldots, N$ i.i.d
2. Set $\widehat{v}_{j+1}^{(n)} = \Psi(v_j^n) + \xi_j^{(n)}$ with $\xi_j^{(n)}$ i.i.d $N(0, \Sigma)$
3. Set $\bar{w}_{j+1}^{(n)} = \exp(-\frac{1}{2}|y_{j+1} - h(\widehat{v}_{j+1}^{(n)})|_\Gamma^2)$
4. Set $w_{j+1}^{(n)} = \bar{w}_{j+1}^{(n)} / \sum_{m=1}^N \bar{w}_{j+1}^{(m)}$
5. Set $\rho_{j+1}^N(u) = \sum_{n=1}^N w_{j+1}^{(n)} \delta(u - \widehat{v}_{j+1}^{(n)})$

3: **Output:** pdf ρ_J^N that approximates the distribution $\mathbb{P}(v_J|Y_J)$

11.3 Bootstrap Particle Filter Convergence

We will now show that under certain conditions, the BPF converges to the true particle filter distribution in the limit $N \to \infty$. The proof is similar to that of the Lax-Equivalence Theorem from the numerical approximation of evolution equations, part of which is the statement that consistency and stability together imply convergence. For the BPF consistency refers to a Monte Carlo error estimate, similar to that derived in the chapter on importance sampling, and stability manifests in bounds on the Lipschitz constants for the operators P and L_j.

Our first step is to define what we mean by convergence, that is, we need a metric on probability measures. Notice that the operators P and L_j are deterministic, but the operator S^N is random since it requires sampling. As a consequence the approximate pdfs ρ_j are also random. Thus, in fact, we need a metric on random probability measures. To this end, for random probability measures ρ and ρ', we define

$$d(\rho, \rho') = \sup_{|f|_\infty \leq 1} \sqrt{\mathbb{E}\,|\rho(f) - \rho'(f)|^2}$$

where the expectation is taken over the random variable, in our case, the randomness from sampling with S^N, and the supreumum is taken over all functions $f : \mathbb{R}^n :\mapsto [-1, 1]$. Here we have used the notation defined in equation (6.2). The following lemma is straightforward to prove, and provides some useful intuition about the metric.

Lemma 11.1. *$d(\cdot, \cdot)$ as defined above does indeed define a metric on random probability measures. Furthermore, when ρ, ρ' are deterministic, then we have $d(\rho, \rho') = 2d_{TV}(\rho, \rho')$.*

We now prove three lemmas which together will enable us to prove convergence of the BPF. The first shows consistency; the second and third show stability estimates for P and L_j respectively.

Lemma 11.2. *Let $\mathcal{P}(\mathbb{R}^n)$ be the set of probability densities on $\mathbb{R}^n$ then*

$$\sup_{\rho\in\mathcal{P}(\mathbb{R}^n)} d(\rho, S^N\rho) \leq \frac{1}{\sqrt{N}}.$$

Proof. First we note

$$S^N\rho(f) = \frac{1}{N}\sum_{j=1}^{N} f(v^{(n)})$$

where $v^{(n)} \sim \rho$ are i.i.d samples. Then

$$\begin{aligned} S^N\rho(f) - \rho(f) &= \frac{1}{N}\sum_{n=1}^{N} f(v^{(n)}) - \rho(f) \\ &= \frac{1}{N}\sum_{n=1}^{N} \bar{f}(v^{(n)}) \end{aligned}$$

with $\bar{f} = f - \rho(f)$. By independence of the samples, it follows that

$$\mathbb{E}\,\bar{f}(v^{(n)})\bar{f}(v^{(l)}) = \delta_{nl}\,\mathbb{E}\,|\bar{f}(v^{(n)})|^2.$$

This is because

$$\mathbb{E}(f(v^{(n)}) - \rho(f)) = \mathbb{E}_{v^{(n)}\sim\rho}\, f(v^{(n)}) - \mathbb{E}_\rho\, f = \mathbb{E}_\rho\, f - \mathbb{E}_\rho\, f = 0.$$

By the same reasoning,

$$\mathbb{E}\,|\bar{f}(v^{(n)})|^2 = \mathbb{E}\,|f(v^{(n)})|^2 - |\,\mathbb{E}\, f(v^{(n)})|^2 \leq 1$$

where the inequality comes from the fact that $|f|_\infty \leq 1$. Hence we find that

$$\mathbb{E}\,|\rho(f) - S^N\rho(f)|^2 = \frac{1}{N^2}\sum_{n=1}^{N} \mathbb{E}\,|\bar{f}(v^{(n)})|^2 \leq \frac{1}{N}.$$

Taking the supremum on both sides gives the desired result. □

Now we prove a stability bound for P.

Lemma 11.3.

$$d(P\mu, P\nu) \leq d(\mu, \nu).$$

Proof. For Markov kernel $p(v', v)$, define function q on $\mathbb{R}^n$ by

$$q(v') = \int_{\mathbb{R}^n} p(v', v) f(v) dv = \mathbb{E}[f(v)|v_0 = v']$$

so

$$|q(v')| \leq \int_{\mathbb{R}^n} p(v', v) dv = 1.$$

Therefore,

$$\begin{aligned}\rho(q) &= \int_{\mathbb{R}^n} q(v')\rho(v')dv' = \int_{\mathbb{R}^n}\left[\int_{\mathbb{R}^n} p(v',v)f(v)dv\right]\rho(v')dv' \\ &= \int_{\mathbb{R}^n}\left[\int_{\mathbb{R}^n} p(v',v)\rho(v')dv'\right]f(v)dv \\ &= \int_{\mathbb{R}^n}(P\rho)(v)f(v)dv\end{aligned}$$

by exchanging the order of integration. Consequently, we have

$$\mathbb{E}^{\rho}[q] = \mathbb{E}^{P\rho}[f].$$

Finally,

$$\begin{aligned}d(P\rho, P\rho') &= \sup_{|f|_\infty \le 1} (\mathbb{E}[|(P\rho)(f) - (P\rho')(f)|^2])^{\frac{1}{2}} \\ &\le \sup_{|q|_\infty \le 1} (\mathbb{E}[|\rho(q) - \rho'(q)|^2])^{\frac{1}{2}} \\ &= d(\rho, \rho').\end{aligned}$$

□

To prove the next lemma, and hence to prove the main convergence theorem about the BPF, we will make the following assumption which encodes the idea of a bound on the observation operator:

Assumption 11.4. *There exists $\kappa \in (0,1)$ such that*

$$\kappa \le g_j(v) \le \kappa^{-1} \quad \forall v \in \mathbb{R}^n, j \in \{1,\dots,J\}.$$

It may initially appear strange to use the same constant κ in the upper and lower bounds, but recall that g is undefined upto a multiplicative constant. Consequently, given any upper and lower bounds, g can be scaled to achieve the bound as stated. Relatedly it is κ^{-2} which appears in the stability constant in the next lemma; if g is not scaled to produce the same constant κ in the upper and lower bounds in Assumption 11.4, then it is the ratio of the upper and lower bounds which would appear in the stability bound.

Lemma 11.5. *Let Assumption 11.4 hold. Then*

$$d(L_j\mu, L_j\nu) \le \frac{2}{\kappa^2} d(\mu,\nu).$$

Proof.

$$\begin{aligned}(L\rho)(f) - (L\rho')(f) &= \frac{\rho(fg)}{\rho(g)} - \frac{\rho'(fg)}{\rho'(g)} \\ &= \frac{\rho(fg)}{\rho(g)} - \frac{\rho'(fg)}{\rho(g)} + \frac{\rho'(fg)}{\rho(g)} - \frac{\rho'(fg)}{\rho(g)'} \\ &= \frac{1}{\kappa}\Big(\frac{\rho(\kappa fg) - \rho'(\kappa fg)}{\rho(g)} + \frac{\rho'(fg)}{\rho'(g)}\frac{\rho'(\kappa g) - \rho(\kappa g)}{\rho(g)}\Big)\end{aligned}$$

Applying Bayes Theorem, we obtain

$$\Big|\frac{\rho'(fg)}{\rho'(g)}\Big| = |\,\mathbb{E}^{L\rho'}(f)| \leq 1.$$

Therefore,

$$|(L\rho)(f) - (L\rho')(f)| \leq \frac{1}{\kappa^2}(|\rho(\kappa fg) - \rho'(\kappa fg)| + |\rho'(\kappa g) - \rho(\kappa g)|)$$

It follows that

$$\mathbb{E}[|(L\rho)(f) - (L\rho')(f)|^2] \leq \frac{2}{\kappa^2}(\mathbb{E}[|\rho(\kappa fg) - \rho'(\kappa fg)|^2] + \mathbb{E}[|\rho'(\kappa g) - \rho(\kappa g)|^2])$$

Since $|\kappa g| \leq 1$ we find that

$$\sup_{|f|_\infty \leq 1} \mathbb{E}[|(L\rho)(f) - (L\rho')(f)|^2] \leq \frac{4}{\kappa^4} \sup_{|f|_\infty \leq 1} \mathbb{E}[|\rho(f) - \rho'(f)|^2]$$

and hence

$$d(L_j\rho, L_j\rho') \leq \frac{2}{\kappa^2} d(\rho, \rho').$$

□

Theorem 11.6 (Convergence of the BPF)**.** *Let Assumption 11.4 hold. Then there exists a* $C = C(J, \kappa)$ *such that*

$$d(\rho_j, \rho_j^N) \leq \frac{C}{\sqrt{N}} \quad \forall j \in 1, \ldots, J$$

Proof. Let $e_j = d(\rho_j, \rho_j^N)$, then

$$\begin{aligned}e_{j+1} = d(\rho_{j+1}, \rho_{j+1}^N) &= d(L_j P\rho_j, L_j S^N P\rho_j^N) \\ &\leq d(L_j P\rho_j^N, L_j P\rho_j) + d(L_j P\rho_j^N, L_j S^N P\rho_j^N)\end{aligned}$$

by the triangle inequality for metrics. Applying the stability bound for L_j, we have

$$e_{j+1} \leq \frac{2}{\kappa^2}[d(P\rho_j^N, P\rho_j) + d(\pi_j^N, S^N\pi_j^N)]$$

where $\pi_j^N = P\rho_j^N$. By the stability bound for P,

$$d(P\rho_j^N, P\rho_j) \leq d(\rho_j^N, \rho_j)$$

and by the consistency bound for S^N

$$d(\pi_j^N, S^N\pi_j^N) \leq \frac{1}{\sqrt{N}}$$

Therefore,

$$\begin{aligned} e_{j+1} &\leq \frac{2}{\kappa^2}\Big(d(\rho_j, \rho_j^N) + \frac{1}{\sqrt{N}}\Big) \\ &\leq \frac{2}{\kappa^2}\Big(e_j + \frac{1}{\sqrt{N}}\Big) \end{aligned}$$

We let $\lambda = 2/\kappa^2$. Then, recalling the Gronwall inequality, we obtain

$$e_j \leq \lambda^j e_0 + \frac{\lambda}{\sqrt{N}}\frac{1-\lambda^j}{1-\lambda}$$

since $\kappa \in (0, 1]$ hence $\lambda \geq 2$. Recall that $\rho_0^N = \rho_0$ hence $e_0 = 0$ then letting

$$C = \frac{\lambda(1-\lambda^J)}{1-\lambda}$$

completes the proof since $\lambda(1-\lambda^j)/(1-\lambda)$ is increasing in j. □

11.4 The Bootstrap Particle Filter as a Random Dynamical System

It is useful for the purposes of analysis to write the BPF as a random dynamical system for a set of interacting particles $\{v_j^{(n)}\}$. To this end we define a measure, of equally weighted particles, which may be naturally created after the resampling step from $\rho_j^N(u)$. We write this measure as

$$\bar{\rho}_j^N(u) = \frac{1}{N}\sum_{n=1}^{N}\delta(u - v_j^{(n)}) \approx \rho_j^N(u) \approx \rho_j(u).$$

Studying the BPF shows that the map on the particle positions may be written as

$$\{v_j^{(n)}\}_{n=1}^N \mapsto \{v_{j+1}^{(n)}\}_{n=1}^N$$

where the map is defined by

$$\begin{aligned} \widehat{v}_{j+1}^{(n)} &= \Psi(v_j^{(n)}) + \xi_j^{(n)} && \xi_j^{(n)} \sim N(0, \Sigma) \text{ i.i.d} \\ v_{j+1}^{(n)} &= \sum_{m=1}^{N} \mathbb{1}_{I_{j+1}^{(m)}}\Big(r_{j+1}^{(n)}\Big)\widehat{v}_{j+1}^{(m)} && r_{j+1}^{(n)} \sim \text{Uniform}(0,1) \text{ i.i.d.} \end{aligned}$$

Here the supports $I_j^{(m)}$ of the indicator functions have widths given by the weights appearing in $\rho_j^N(u)$. Specifically we have

$$I_{j+1}^{(m)} = \left[\alpha_{j+1}^{(m-1)}, \alpha_{j+1}^{(m)}\right) \qquad \alpha_{j+1}^{(m+1)} = \alpha_{j+1}^{(m)} + w_{j+1}^{(m)}, \quad \alpha_{j+1}^{(0)} = 0.$$

Note that, by construction, $\alpha_j^{(N)} = 1$.

Thus the underlying dynamical system on particles comprises N particles governed by two steps: (i) the underlying stochastic dynamics model, in which the particles do not interact; (ii) a resampling of the resulting collection of particles, to reflect the different weights associated with them, in which the particles do then interact. The interaction is driven by the weights which see all the particle positions, and measure their goodness of fit to the data.

11.5 Discussion and Bibliography

Particle filters are overviewed from an algorithmic viewpoint in [27, 28], and from a more mathematical perspective in [25]. The convergence of particle filters was addressed in [20]; the clean proof presented here originates in [93] and may also be found in [70]. For problems in which the dynamics evolve in relatively low dimensional spaces they have been an enormously successful. Generalizing them so that they work for the high dimensional problems that arise, for example, in geophysical applications, provides a major challenge [109]; the next two chapters are related to methodologies which address this challenge.

12 Optimal Particle Filter

This lecture is devoted to the optimal particle filter. Like the bootstrap filter from the previous lecture this is a method designed to approximate the true filtering distribution by a sum of Dirac measures. The setting will initially be the same as in the previous lecture (nonlinear stochastic dynamics and nonlinear observations). However the optimal particle filter can not, in general, be tractabily implemeneted. However it may be implemented in a straightforward fashion when the observation operator is linear; it may then be characterized as a set of interacting 3DVAR filters. We therefore introduce the concept in the fully nonlinear setting, and then specialize to the case of linear observation operator.

12.1 Introduction

The bootstrap particle filter from the previous lecture is based on approximating the two components of a specific factorization of the particle filtering map. The factorization is

$$\begin{aligned} \mathbb{P}(v_{j+1}|Y_j) &= P\,\mathbb{P}(v_j|Y_j), \\ \mathbb{P}(v_{j+1}|Y_{j+1}) &= L_j\,\mathbb{P}(v_{j+1}|Y_j). \end{aligned}$$

This gives the factorization

$$\mathbb{P}(v_{j+1}|Y_{j+1}) = L_j P\,\mathbb{P}(v_j|Y_j)$$

which is the basis for the bootstrap particle filter. It is natural to ask if there are other factorizations of the filtering update and whether they might lead to improved particle filters. In this lecture we derive the Optimal Particle Filter (OPF) which does just that. We demonstrate a connection with 3DVAR, and we discuss the sense in which the OPF has desirable properties in comparison with the BPF. Throughout we write ρ_j for the probability density $\mathbb{P}(v_j|Y_j)$.

12.2 The Bootstrap and Optimal Particle Filters Compared

The fundamental filtering problem that we are interested in is determination of $\mathbb{P}(v_{j+1}|Y_{j+1})$ from $\mathbb{P}(v_j|Y_j)$. In the BPF, we are approximating the following manipulation:

$$\begin{aligned} \mathbb{P}(v_{j+1}|Y_{j+1}) &= \mathbb{P}(v_{j+1}|y_{j+1}, Y_j) \\ &= \frac{\mathbb{P}(y_{j+1}|v_{j+1}, Y_j)\,\mathbb{P}(v_{j+1}|Y_j)}{\mathbb{P}(y_{j+1}|Y_j)} \qquad \text{from Bayes' rule} \\ &= \int_{\mathbb{R}^N} \frac{\mathbb{P}(y_{j+1}|v_{j+1}, Y_j)\,\mathbb{P}(v_{j+1}|v_j, Y_j)\,\mathbb{P}(v_j|Y_j)}{\mathbb{P}(y_{j+1}|Y_j)} dv_j \\ &= \int_{\mathbb{R}^N} \frac{\mathbb{P}(y_{j+1}|v_{j+1}, Y_j)\,\mathbb{P}(v_{j+1}|v_j)\,\mathbb{P}(v_j|Y_j)}{\mathbb{P}(y_{j+1}|Y_j)} dv_j \\ &= \int_{\mathbb{R}^N} \frac{\mathbb{P}(y_{j+1}|v_{j+1})\,\mathbb{P}(v_{j+1}|v_j)\,\mathbb{P}(v_j|Y_j)}{\mathbb{P}(y_{j+1}|Y_j)} dv_j \\ &= L_j^{BPF} P^{BPF}\,\mathbb{P}(v_j|Y_j) \end{aligned}$$

with Markov kernel for particle update

$$P^{BPF}\nu(v_{j+1}) = \int_{\mathbb{R}^N} \mathbb{P}(v_{j+1}|v_j)\nu(v_j)dv_j$$

and application of Bayes' theorem, taking into account the likelihood of the data

$$L_j^{BPF}\nu(v_{j+1}) = \frac{1}{Z_\nu}\,\mathbb{P}(y_{j+1}|v_{j+1})\nu(v_{j+1}),$$

with Z_ν normalization to a probability density. Thus we have

$$\rho_{j+1} = L_j^{BPF} P_j^{BPF} \rho_j. \tag{12.1}$$

Note that in this factorization we apply a Markov kernel and then Bayes rule. In contrast, in the OPF, we begin with the same expression but perform different manipulations:

$$\begin{aligned}
\mathbb{P}(v_{j+1}|Y_{j+1}) &= \int_{\mathbb{R}^N} \mathbb{P}(v_{j+1}, v_j|Y_{j+1})dv_j \\
&= \int_{\mathbb{R}^N} \mathbb{P}(v_{j+1}|v_j, Y_{j+1})\,\mathbb{P}(v_j|Y_{j+1})dv_j \text{ (conditional expectation, } Y_j \text{ fixed)} \\
&= \int_{\mathbb{R}^N} \mathbb{P}(v_{j+1}|v_j, y_{j+1}, Y_j)\,\mathbb{P}(v_j|y_{j+1}, Y_j)dv_j \\
&= \int_{\mathbb{R}^N} \mathbb{P}(v_{j+1}|v_j, y_{j+1})\,\mathbb{P}(v_j|y_{j+1}, Y_j)dv_j \\
&= \int_{\mathbb{R}^N} \frac{\mathbb{P}(v_{j+1}|v_j, y_{j+1})\,\mathbb{P}(y_{j+1}|v_j, Y_j)\,\mathbb{P}(v_j|Y_j)}{\mathbb{P}(y_{j+1}|Y_j)}dv_j \quad \text{(Bayes' Rule)} \\
&= \int_{\mathbb{R}^N} \frac{\mathbb{P}(v_{j+1}|v_j, y_{j+1})\,\mathbb{P}(y_{j+1}|v_j)\,\mathbb{P}(v_j|Y_j)}{\mathbb{P}(y_{j+1}|Y_j)}dv_j \\
&= Q_j^{OPF} L_j^{OPF}\,\mathbb{P}(v_j, Y_j)
\end{aligned}$$

with Markov kernel for particle update

$$Q_j^{OPF}\nu(v_{j+1}) = \int_{\mathbb{R}^N} \mathbb{P}(v_{j+1}|v_j, y_{j+1})\nu(v_j)dv_j$$

and application of Bayes' rule to include the likelihood

$$L_j^{OPF}\nu(v_j) = \frac{1}{Z_\nu}\,\mathbb{P}(y_{j+1}|v_j)\nu(v_j).$$

Thus we have

$$\rho_{j+1} = Q_j^{OPF} L_j^{OPF} \rho_j. \tag{12.2}$$

Note that in this factorization we apply Bayes rule and then a Markov kernel, the opposite order to the BPF. However, once particle approximations are introduced this distinction disappears, but we are left with a different propagation mechanism for the particles – one that sees the data through the Markov kernel Q_j^{OPF} – and hence a different weighting of the particles. We will see that the particle updates use a 3DVAR procedure. In the BPF, the evolution of the particles and the observation of the data are kept separate from each other – the Markov kernel P^{BPF} depends only on the dynamics and not the observed data and is thus independent of j.

12.3 Implementation of Optimal Particle Filter: Linear Observation Operator

In general it is not possible to implement the OPF because exact sampling from the Markov kernel is not possible. However if we assume that the observation function $h(\cdot)$ is linear, i.e. $h(\cdot) = H$ then exact sampling is possible and so we concentrate on this setting; it arises in many applications. Given this, the dynamics/data setting is given by

$$\begin{aligned} v_{j+1} &= \Psi(v_j) + \xi_j & \xi_j &\sim N(0, \Sigma) \text{ i.i.d} \\ y_{j+1} &= Hv_{j+1} + \eta_{j+1} & \eta_{j+1} &\sim N(0, \Gamma) \text{ i.i.d} \end{aligned}$$

with $v_0 \sim N(m_0, C_0)$ and $v_0, \{\xi_j\}, \{\eta_j\}$ independent. Combining the dynamics and data models we may therefore write

$$y_{j+1} = H\Psi(v_j) + H\xi_j + \eta_{j+1}$$

which gives us a conditional distribution for y_{j+1}:

$$\mathbb{P}(y_{j+1}|v_j) = N(H\Psi(v_j), S).$$

where $S = H\Sigma H^T + \Gamma$. We will use this formula to implement L_j^{OPF}.

We now determine Q_j^{OPF}. Looking now to the posterior distribution for v_{j+1}, we have

$$\begin{aligned} \mathbb{P}(v_{j+1}|v_j, y_{j+1}) &\propto \mathbb{P}(y_{j+1}|v_{j+1}, v_j)\,\mathbb{P}(v_{j+1}|v_j) \\ &= \mathbb{P}(y_{j+1}|v_{j+1})\,\mathbb{P}(v_{j+1}|v_j) \\ &\propto \exp\left(-\frac{1}{2}|y_{j+1} - Hv_{j+1}|_\Gamma^2 - \frac{1}{2}|v_{j+1} - \Psi(v_j)|_\Sigma^2\right) \\ &= \exp(-\mathsf{J}_{\text{OPT}}(v_{j+1})). \end{aligned}$$

This is a Gaussian distribution for v_{j+1} as $\mathsf{J}_{\text{OPT}}(v_{j+1})$ is quadratic in respect to v_{j+1}.[4] Consequently, we can compute the mean m_{j+1} and covariance C (which, note, is independent of j) of this Gaussian by matching the mean and quadratic terms in the relevant quadratic forms:

$$\begin{aligned} C^{-1} &= H^T\Gamma^{-1}H + \Sigma^{-1} \\ C^{-1}m_{j+1} &= \Sigma^{-1}\Psi(v_j) + H^T\Gamma^{-1}y_{j+1} \end{aligned}$$

Then $\mathbb{P}(v_{j+1}|y_{j+1}, v_j) = N(m_{j+1}, C)$. This is hence a special case of 3DVAR in which the analysis covariance is fixed at C; note that when we derived 3DVAR we fixed the predictive covariance $\widehat{C}$ which, here, is fixed at Σ. As with the Kalman filter, and with 3DVAR, it is possible to implement the prediction step through the following mean and

[4] J_{OPT} is identical to J on the right-hand side of Table 10.1, with $\widehat{C}$ replaced by Σ.

covariance formulae which avoid inversion in state space, and require inversion only in data space:

$$\begin{aligned} m_{j+1} &= (I - KH)\Psi(v_j) + Ky_{j+1} \\ C &= (I - KH)\Sigma \\ K &= \Sigma H^T S^{-1} \\ S &= H\Sigma H^T + \Gamma \end{aligned}$$

Furthermore, as for 3DVAR, the inversion of S need only be performed once in a pre-processing step before the algorithm is run. Since the expression for $\mathbb{P}(v_{j+1}|v_j, y_{j+1})$ is Gaussian we now have the ability to sample directly from Q_j^{OPF}. The OPF is thus given by the following update algorithm for approximations $\rho_j^N \approx \rho_j$ in which we generalize the notational conventions used in the previous lecture to formulate particle filters as random dynamical systems:

Algorithm 12.1 Algorithm for the Optimal Particle Filter with linear observation map H

1: **Input**: Initial distribution $\mathbb{P}(v_0) = \rho_0$, observations Y_J, number of particles N
2: **Initial Sampling**: Draw N particles $v_0^{(n)} \sim \rho_0$ so that $\rho_0^N = S^N \rho_0$
3: **Subsequent Sampling** For $j = 0, 1, \ldots, J-1$, perform

1. Set $\widehat{v}_{j+1}^{(n)} = (I - KH)\Psi(v_j^{(n)}) + Ky_{j+1} + \zeta_{j+1}^{(n)}$ with $\zeta_{j+1}^{(n)}$ i.i.d $N(0, C)$
2. Set $\bar{w}_{j+1}^{(n)} = \exp\Big(-\frac{1}{2}|y_{j+1} - H\Psi(v_j^{(n)})|_S^2\Big)$
3. Set $w_{j+1}^{(n)} = \bar{w}_{j+1}^{(n)} / \sum_{i=1}^N \bar{w}_{j+1}^{(n)}$
4. Set $v_{j+1}^{(n)} = \sum_{m=1}^N \mathbb{1}_{I_{j+1}^{(m)}}(r_{j+1}^{(n)}) \widehat{v}_{j+1}^{(m)}$
5. Set $\rho_{j+1}^N(u) = \frac{1}{N} \sum_{n=1}^N \delta\Big(u - v_{j+1}^{(n)}\Big)$

4: **Output**: N particles $v_J^1, v_J^2, \ldots, v_J^N$

It would be desirable to interpret this algorithm as an approximation of the filter update (12.2) in the form

$$\rho_{j+1}^N = S^N Q_j^{OPF} L_j^{OPF} \rho_j^N, \quad \rho_j^0 = S^N \rho_0.$$

However the order in which the resampling and the particle propagation occurs means that this is not possible. The following slight modification of the OPF, however, may indeed be thought of as an approximation of this form; we simply reorder the resampling and the propagation. We refer to the resulting algorithm as the Gaussianized Optimal Particle filter (GOPF). We may write the resulting algorithm as follows:

Algorithm 12.2 Algorithm for the Gaussianized Optimal Particle Filter

1: **Input**: Initial distribution $\mathbb{P}(v_0) = \rho_0$, observations Y_J, number of particles N
2: **Initial Sampling**: Draw N particles $v_0^{(n)} \sim \rho_0$ so that $\rho_0^N = S^N \rho_0$
3: **Subsequent Sampling** For $j = 0, 1, \ldots, J-1$, perform

1. Set $\bar{w}_{j+1}^{(n)} = \exp\Big(-\frac{1}{2}|y_{j+1} - H\Psi(v_j^{(n)})|_S^2\Big)$
2. Set $w_{j+1}^{(n)} = \bar{w}_{j+1}^{(n)} / \sum_{i=1}^N \bar{w}_{j+1}^{(n)}$
3. Set $\widehat{v}_j^{(n)} = \sum_{m=1}^N \mathbb{1}_{I_{j+1}^{(m)}}(r_{j+1}^{(n)}) v_j^{(m)}$
4. Set $v_{j+1}^{(n)} = (I - KH)\Psi(\widehat{v}_j^{(n)}) + K y_{j+1} + \zeta_{j+1}^{(n)}$ with $\zeta_{j+1}^{(n)}$ i.i.d $N(0, C)$
5. Set $\rho_{j+1}^N(u) = \frac{1}{N} \sum_{n=1}^N \delta\Big(u - v_{j+1}^{(n)}\Big)$

4: **Output**: N particles $v_J^1, v_J^2, \ldots, v_J^N$

12.4 "Optimality" of the Optimal Particle Filter

Particle filter methods rely on approximating the distribution for the model v by a swarm of point Dirac functions; it is clear that the distribution will not be well approximated by only a small number of particles in most cases. Consequently, a performance requirement for particle filter methods is that they do not lead to degeneracy of the particles. Resampling leads to degeneracy if a small number of particles have all the weights. Conversely, non-degeneracy may be promoted by ensuring that the weights $w_j^{(n)}$ are similar in magnitude, so that a small number of particles are not overly favoured during the resampling step. This condition can be formulated as a requirement that the variance of the weights be minimized; doing this results in the OPF.

To understand this perspective we consider an arbitrary particle update kernel of the form $\pi(v_{j+1}|v_j^{(n)}, Y_{j+1})$ and we study the resulting particle filter without resampling. It is then the case that the particle weights are updated according to the formula

$$\bar{w}_{j+1}^{(n)} = \bar{w}_j^{(n)} \frac{\mathbb{P}(y_{j+1}|v_{j+1})\, \mathbb{P}(v_{j+1}|v_j^{(n)})}{\pi(v_{j+1}|v_j^{(n)}, Y_{j+1})}.$$

Theorem 12.1 (Meaning of Optimality). *The choice of $\mathbb{P}(v_{j+1}|v_j^{(n)}, y_{j+1})$ as the particle update kernel $\pi(v_{j+1}|v_j^{(n)}, Y_{j+1})$ results in the minimal variance of the weight $w_{j+1}^{(n)}$, with respect to all possible choices of the particle update kernel $\pi(v_{j+1}|v_j^{(n)}, Y_{j+1})$.*

Proof. We calculate the variance of the unnormalized weights (treated as random

variables) $\bar{w}_{j+1}^{(n)}$ with respect to the transition density $\pi(v_{j+1}|v_j^{(n)}, Y_{j+1})$ and obtain

$$\begin{aligned}
\mathrm{Var}_{\pi(v_{j+1}|v_j,y_{j+1})}[\bar{w}_{j+1}^{(n)}] &= \int_{\mathbb{R}^N} \left(\bar{w}_{j+1}^{(n)}\right)^2 \pi(v_{j+1}|v_j^{(n)}, Y_{j+1}) dv_{j+1} \\
&- \left[\int_{\mathbb{R}^N} \bar{w}_{j+1}^{(n)} \pi(v_{j+1}|v_j^{(n)}, Y_{j+1}) dv_{j+1}\right]^2 \\
&= \left(\bar{w}_j^{(n)}\right)^2 \int_{\mathbb{R}^N} \frac{\left(\mathbb{P}(y_{j+1}|v_{j+1})\,\mathbb{P}(v_{j+1}|v_j^{(n)})\right)^2}{\pi(v_{j+1}|v_j^{(n)}, Y_{j+1})} dv_{j+1} \\
&\qquad - \left(\bar{w}_j^{(n)}\right)^2 \left[\int_{\mathbb{R}^N} \mathbb{P}(y_{j+1}|v_{j+1})\,\mathbb{P}(v_{j+1}|v_j^{(n)}) dv_{j+1}\right]^2 \\
&= \left(\bar{w}_j^{(n)}\right)^2 \left[\int_{\mathbb{R}^N} \frac{\left(\mathbb{P}(y_{j+1}|v_{j+1})\,\mathbb{P}(v_{j+1}|v_j^{(n)})\right)^2}{\pi(v_{j+1}|v_j^{(n)}, Y_{j+1})} dv_{j+1} - \mathbb{P}(y_{j+1}|v_j^{(n)})^2\right]
\end{aligned}$$

Choosing $\pi(v_{j+1}|v_j, Y_{j+1}) = \mathbb{P}(v_{j+1}|v_j, y_{j+1})$, as in the OPF, we obtain

$$\begin{aligned}
\mathrm{Var}_{\mathbb{P}(v_{j+1}|v_j,y_{j+1})}[\bar{w}_{j+1}^{(n)}] &= \left(\bar{w}_j^{(n)}\right)^2 \left[\int_{\mathbb{R}^N} \frac{\left(\mathbb{P}(y_{j+1}|v_{j+1})\,\mathbb{P}(v_{j+1}|v_j^{(n)})\right)^2}{\mathbb{P}(v_{j+1}|v_j^{(n)}, y_{j+1})} dv_{j+1} - \mathbb{P}(y_{j+1}|v_j^{(n)})^2\right] \\
&= \left(\bar{w}_j^{(n)}\right)^2 \left[\mathbb{P}(y_{j+1}|v_j^{(n)})^2 - \mathbb{P}(y_{j+1}|v_j^{(n)})^2\right] \\
&= 0.
\end{aligned}$$

□

Remark 12.2. The optimal particle filter is optimal in the very precise sense of the theorem. Note in particular that no optimality criterion is asserted by this theorem with respect to iterating the particle updates, and in particular when resampling is included. The nomenclature "optimal" should thus be treated with caution.

12.5 Particle Filters for High Dimensions

Particle filters often perform poorly for high-dimensional systems due to a collapse of the particle weights: only a few particles carry most of the weight; the optimal particle filter can ameliorate this issue because the proposal uses the data, meaning that particle predictions are more likely to be weighted highly. There have been attempts to formulate update steps that help to mitigate this weight collapse (see [101]). Essentially, these methods aim to push the particles towards the region of high likelihood, such that all the particles will be representative of the distribution. As will show in the next lecture, there are interesting particle based methods which, whilst not statistically consistent, do perform well as signal estimators. This perspective on particle filter type methods, namely to use them for smart signal estimation rather than as estimators of the filtering distribution, may become increasingly useful in high dimensional systems.

12.6 Discussion and Bibliography

Particle filters often perform poorly for high-dimensional systems due to the fact that the particle weight typically concentrates on one, or a small number, of particles –see the work of Bickel and Snyder in [9, 100, 101]. This is the issue that the optimal particle filter tries to ameliorate; the paper [4] shows calculations which demonstrate the extent to which this amelioration is manifest in theory. The optimal particle filter is discussed, and further references given, in the very clear paper [28]; see section IID. Throughout much of this lecture we consider the case of Gaussian additive noise and linear observation operator in which case the prediction step is tractable; the paper [28] discusses the more general setting. The order in which the prediction and resampling is performed can be commuted in this case and a discussion of this fact may be found in [91]; this leads to the distinction between what we term the GOPF and the OPF. The convergence of the optimal particle filter is studied in [56]. The formulation of the bootstrap and optimal particle filters as random dyamical systems may be found in [64].

13 The Extended and Ensemble Kalman Filters

In this chapter, the Extended Kalman Filter (ExKF)[5] and the Ensemble Kalman Filter (EnKF) are described. The status of the two methods in relation to the true filtering distribution is as follows. The ExKF is a provably accurate approximation of the true filtering distribution in situations in which small noise is present in both signal and data, and the filtering distribution is well-approximated by a Gaussian. The EnKF is also in principle a good approximation of the filtering distribution in this situation, if a large number of particles is used. However the EnKF is typically deployed for problems where the use of a sufficiently large number of particles is impractical; it is then better viewed as an online optimizer, in the spirit of 3DVAR, but using multiple particles to better estimate the covariances appearing in the quadratic objective function which is minimized to find particle updates. For clarity we conclude this chapter by summarizing all of the filtering methods introduced in this and preceding chapters. We focus on summarizing their differences and the setting in which they may be practically applied.

13.1 The Setting

Throughout this chapter we consider the setting in which 3DVAR was introduced and may be applied: the dynamics model is nonlinear, but the observation operator is linear. For purposes of exposition we summarize it again here:

$$\begin{aligned} v_{j+1} &= \Psi(v_j) + \xi_j & \xi_j &\sim N(0,\Sigma) \text{ i.i.d.}, \\ y_{j+1} &= Hv_{j+1} + \eta_j & \eta_j &\sim N(0,\Gamma) \text{ i.i.d.}, \end{aligned}$$

with $v_0 \sim N(m_0, C_0)$ independent of the independent i.i.d. sequences $\{\xi_j\}$ and $\{\eta_j\}$. Throughout this chapter we assume that $v_j \in \mathbb{R}^\ell, y_j \in \mathbb{R}^m$ and $H \in \mathbb{R}^{m\times\ell}$. The reason for using ℓ rather than n is to avoid notational conflict with n which here indexes ensemble members.

13.2 The Extended Kalman Filter

This method is derived by applying the Kalman methodology, using linearization to propagate the covariance C_j to the predictive covariance $\widehat{C}_{j+1}$. The Table 13.1 summarizes the idea, after which we calculate the formulae required in full detail.

[5] The extended Kalan filter is often termed the EKF in the literature, a terminology introduced before the existence of the EnKF; we find it useful to write ExKF to uneqivocally distinguish it from the EnKF.

Kalman Filter	ExKF
$m_{j+1} = \arg\min_m \mathsf{J}(m)$	$m_{j+1} = \arg\min_m \mathsf{J}(m)$
$\mathsf{J}(m) = \frac{1}{2}\lvert y_{j+1} - Hm\rvert^2_\Gamma + \frac{1}{2}\lvert m - \widehat{m}_{j+1}\rvert^2_{\widehat{C}_{j+1}}$	$\mathsf{J}(m) = \frac{1}{2}\lvert y_{j+1} - Hm\rvert^2_\Gamma + \frac{1}{2}\lvert m - \widehat{m}_{j+1}\rvert^2_{\widehat{C}_{j+1}}$
$\widehat{m}_{j+1} = Mm_j$	$\widehat{m}_{j+1} = \Psi(m_j)$
$\widehat{C}_{j+1}$ update exact	$\widehat{C}_{j+1}$ update by linearization
$m_{j+1} = (I - K_{j+1}H)\widehat{m}_{j+1} + K_{j+1}y_{j+1}$	$m_{j+1} = (I - K_{j+1}H)\widehat{m}_{j+1} + K_{j+1}y_{j+1}$

Table 13.1 Comparison of Kalman Filter and ExKF update formulae

We first recall the Kalman filter update formulae and their derivation. We have

$$\widehat{v}_{j+1} = Mv_j + \xi_j, \quad v_j \sim N(m_j, C_j), \quad \xi_j \sim N(0, \Sigma) \tag{13.1}$$

From this we deduce, by taking expectations, that

$$\widehat{m}_{j+1} = \mathbb{E}(\widehat{v}_{j+1} \mid Y_j) = \mathbb{E}(Mv_j + \xi_j \mid Y_j) = \mathbb{E}(Mv_j \mid Y_j) + \mathbb{E}(\xi_j \mid Y_j) = Mm_j \tag{13.2}$$

The covariance update is derived as follows.

$$\begin{aligned}
\widehat{C}_{j+1} &= \mathbb{E}((\widehat{v}_{j+1} - \widehat{m}_{j+1}) \otimes (\widehat{v}_{j+1} - \widehat{m}_{j+1}) \mid Y_j) \\
&= \mathbb{E}((M(v_j - m_j) + \xi_j) \otimes (M(v_j - m_j) + \xi_j) \mid Y_j) \\
&= \mathbb{E}((M(v_j - m_j)) \otimes (M(v_j - m_j)) \mid Y_j) + \mathbb{E}(\xi_j \otimes \xi_j \mid Y_j) \\
&\quad + \mathbb{E}((M(v_j - m_j)) \otimes \xi_j \mid Y_j) + \mathbb{E}(\xi_j \otimes (M(v_j - m_j)) \mid Y_j) \\
&= M\,\mathbb{E}((v_j - m_j) \otimes (v_j - m_j) \mid Y_j)M^T + \Sigma \\
&= MC_jM^T + \Sigma.
\end{aligned} \tag{13.3}$$

For the ExKF, the prediction map Ψ is no longer linear. But since ξ_j is independent of Y_j and v_j, we obtain

$$\widehat{m}_{j+1} = \mathbb{E}(\Psi(v_j) + \xi_j \mid Y_j) = \mathbb{E}(\Psi(v_j) \mid Y_j) + \mathbb{E}(\xi_j \mid Y_j) = \mathbb{E}(\Psi(v_j) \mid Y_j).$$

If we assume that the fluctuations of v_j around its mean m_j (conditional on data) are small then a reasonable approximation is to take $\Psi(v_j) \approx \Psi(m_j)$ so that

$$\widehat{m}_{j+1} = \Psi(m_j). \tag{13.4}$$

For the predictive covariance we use linearization; we have

$$\begin{aligned}
\widehat{C}_{j+1} &= \mathbb{E}((\widehat{v}_{j+1} - \widehat{m}_{j+1}) \otimes (\widehat{v}_{j+1} - \widehat{m}_{j+1}) \mid Y_j) \\
&= \mathbb{E}((\Psi(v_j) - \Psi(m_j) + \xi_j) \otimes (\Psi(v_j) - \Psi(m_j) + \xi_j) \mid Y_j) \\
&= \mathbb{E}((\Psi(v_j) - \Psi(m_j)) \otimes (\Psi(v_j) - \Psi(m_j) \mid Y_j) + \Sigma \\
&\approx D\Psi(m_j)\,\mathbb{E}((v_j - m_j) \otimes (v_j - m_j) \mid Y_j)D\Psi(m_j)^T + \Sigma
\end{aligned}$$

and so, again assuming that fluctuations of v_j around its mean m_j (conditional on data) are small we invoke the approximation

$$\widehat{C}_{j+1} = D\Psi(m_j)C_j D\Psi(m_j)^T + \Sigma. \tag{13.5}$$

To be self-consistent Σ itself should be small.

The analysis step is the same as for the Kalman filter:

$$S_{j+1} = H\widehat{C}_{j+1}H^T + \Gamma \tag{13.6a}$$
$$K_{j+1} = \widehat{C}_{j+1}H^T S_{j+1}^{-1} \qquad \text{(Gain Matrix)} \tag{13.6b}$$
$$m_{j+1} = (I - K_{j+1}H)\widehat{m}_{j+1} + K_{j+1}y_{j+1} \tag{13.6c}$$
$$C_{j+1} = (I - K_{j+1}H)\widehat{C}_{j+1} \tag{13.6d}$$
$$\tag{13.6e}$$

Thus the overall ExKF comprises equations (13.4), (13.5) and (13.6). Unlike the Kalman filter, for the extended Kalman filter the maps $C_j \mapsto \widehat{C}_{j+1} \mapsto C_{j+1}$ depend on the observed data, through the dependence of the predictive covariance on the filter mean. To be self-consistent with the "small fluctuations around the mean" assumptions made in the derivation of the ExKF, Σ and Γ should both be small.

The analysis step can also be defined by

$$C_{j+1}^{-1} = \widehat{C}_{j+1}^{-1} + H^T\Gamma^{-1}H \tag{13.7a}$$
$$m_{j+1} = \operatorname*{argmin}_v \mathsf{J}(v) \tag{13.7b}$$

where

$$\mathsf{J}(v) = \frac{1}{2} \mid y_{j+1} - Hv \mid_{\Gamma}^2 + \frac{1}{2} \mid v - \widehat{m}_{j+1} \mid_{\widehat{C}_{j+1}}^2 \tag{13.7c}$$

and $\widehat{m}_{j+1}, \widehat{C}_{j+1}$ are calculated as above in the prediction steps (13.4), (13.5). The constraint formulation of the minimization problem, derived for the Kalman filter in section 9.4, may also be used to derive the update formulae above.

13.3 Ensemble Kalman Filter

When the dynamical system is in high dimension, evaluation and storage of the predicted covariance, and in particular the Jacobian required for the update formula (13.5), becomes computationally inefficient and expensive for the ExKF. The EnKF was developed to overcome this issue. The basic idea is to maintain an ensemble of particles, as in the particle filters, and to use their empirical covariance within a Kalman-type update. The method is summarized in the Table 13.2. It may be thought of as an ensemble 3DVAR technique in which a colletion of particles are generated similarly to 3DVAR, but interact through an ensemble estimate of their covariance.

Kalman Filter	EnKF
$m_{j+1} = \arg\min_m J(m)$	$m_{j+1} = \arg\min_m J(m)$
$J(m) = \frac{1}{2}\lvert y_{j+1} - Hm\rvert^2_\Gamma + \frac{1}{2}\lvert m - \widehat{m}_{j+1}\rvert^2_{\widehat{C}_{j+1}}$	$J_n(m) = \frac{1}{2}\lvert y^{(n)}_{j+1} - Hm\rvert^2_\Gamma + \frac{1}{2}\lvert m - \widehat{v}^{(n)}_{j+1}\rvert^2_{\widehat{C}_{j+1}}$
$\widehat{m}_{j+1} = Mm_j$	$\widehat{v}^{(n)}_{j+1} = \Psi(v^{(n)}_j) + \xi^{(n)}_j$
$\widehat{C}_{j+1}$ update exact	$\widehat{C}_{j+1}$ update by ensemble estimate
$m_{j+1} = (I - K_{j+1}H)\widehat{m}_{j+1} + K_{j+1}y_{j+1}$	$m_{j+1} = (I - K_{j+1}H)\widehat{m}_{j+1} + K_{j+1}y_{j+1}$

Table 13.2 Comparison of Kalman Filter and EnKF update formulae

13.4 Derviation of EnKF

In the basic form which we present here, the EnKF is applied when Ψ is nonlinear, while the observation operator H is linear. The N particles used at step j are denoted $\{v^{(n)}_j\}_{n=1}^N$. They are all given equal weight so that it is possible to think of making an approximation to the filtering distribution of the form

$$\rho^N_j \approx \frac{1}{N}\sum_{n=1}^N \delta(u - v^{(n)}_j).$$

However the EnKF is typically used with a relatively small number N of particles and may be far from approximating the desired distribution in $\mathbb{R}^\ell$. It is then better understood as a sequential optimization method, similar in spirit to 3DVAR, as described above; this is our perspective.

The state of all the particles at time $j+1$ are predicted to give $\{\widehat{v}^{(n)}_{j+1}\}_{n=1}^N$ using the dynamical model. The resulting empirical covariance is then used to define the objective function (13.7c) which is minimized in order to perform the analysis step and obtain $\{v^{(n)}_{j+1}\}_{n=1}^N$. The updates are denoted schematically by

$$\{v^{(n)}_j\}_{n=1}^N \overset{\text{ⓟ}}{\longmapsto} \{\widehat{v}^{(n)}_{j+1}\}_{n=1}^N \overset{\text{ⓐ}}{\longmapsto} \{v^{(n)}_{j+1}\}_{n=1}^N$$

We now detail these two steps.

ⓟ prediction:

$$\widehat{v}^{(n)}_{j+1} = \Psi(v^{(n)}_j) + \xi^{(n)}_j, n = 1, ..., N \tag{13.8a}$$

$$\widehat{m}_{j+1} = \frac{1}{N}\sum_{n=1}^N \widehat{v}^{(n)}_{j+1} \tag{13.8b}$$

$$\widehat{C}_{j+1} = \frac{1}{N}\sum_{n=1}^N (\widehat{v}^{(n)}_{j+1} - \widehat{m}_{j+1}) \otimes (\widehat{v}^{(n)}_{j+1} - \widehat{m}_{j+1}) \tag{13.8c}$$

Here we have

$$\xi^{(n)}_j \sim N(0, \Sigma), i.i.d$$

ⓐ analysis:

$$S_{j+1} = H\widehat{C}_{j+1}H^T + \Gamma \tag{13.9a}$$

$$K_{j+1} = \widehat{C}_{j+1}H^T S_{j+1}^{-1} \qquad \text{(Gain Matrix)} \tag{13.9b}$$

$$y_{j+1}^{(n)} = y_{j+1} + s\eta_{j+1}^{(n)}, \quad n = 1, ..., N \tag{13.9c}$$

$$v_{j+1}^{(n)} = (I - K_{j+1}H)\widehat{v}_{j+1}^{(n)} + K_{j+1}y_{j+1}^{(n)}, \quad n = 1, ..., N \tag{13.9d}$$

$$\tag{13.9e}$$

Here we take

$$\eta_j^{(n)} \sim N(0, \Gamma), i.i.d$$

The constant s takes value 0 or 1. When $s = 1$ the $y_{j+1}^{(n)}$ are referred to as *perturbed observations*. The analysis step may be written as

$$v_{j+1}^{(n)} = \underset{v}{\text{argmin}}\, \mathsf{J}_n(v) \tag{13.10}$$

where

$$\mathsf{J}_n(v) := \frac{1}{2} \mid y_{j+1}^{(n)} - Hv \mid_{\Gamma}^2 + \frac{1}{2} \mid v - \widehat{v}_{j+1}^{(n)} \mid_{\widehat{C}_{j+1}}^2 \tag{13.11}$$

and the predicted mean and covariance are given by step ⓓ. Note that $\widehat{C}_{j+1}$ is typically not invertible as it is a rank N matrix and N is usually less than the dimension ℓ of the space on which $\widehat{C}_{j+1}$ acts; this is since the typical use of ensemble methods is for high dimensional state space estimation, with a small ensemble size. The minimizing solution can be found by regularizing $\widehat{C}_{j+1}$ by additions of ϵI for $\epsilon > 0$, deriving the update equations as above for ⓐ, and then letting $\epsilon \to 0$. Alternatively, the constraint formulation of the minimization problem, derived for the Kalman filter in section 9.4, may also be used to derive the update formulae above.

13.5 Subspace Property of EnKF

We now give another way to think of, and exploit in algorithms, the low rank property of $\widehat{C}_{j+1}$. Note that $J_n(v)$ is undefined unless

$$v - \widehat{v}_{j+1}^{(n)} = \widehat{C}_{j+1}a$$

for some $a \in \mathbb{R}^\ell$. From the structure of $\widehat{C}_{j+1}$ given in (13.8c) it follows that

$$v = \widehat{v}_{j+1}^{(n)} + \frac{1}{N}\sum_{m=1}^{N} b_m(\widehat{v}_{j+1}^{(m)} - \widehat{m}_{j+1}) \tag{13.12}$$

for some unknown parameters $\{b_m\}_{m=1}^N$, to be determined. (Note that the vector $\{b_m\}_{m=1}^N$ depends on enesemble index n; we have supressed this dependence for notational convenience). This form for v can be substituted into (13.11) to obtain a functional $I_n(b)$ to be minimized over $b \in \mathbb{R}^N$. We re-emphasize that N will typically

be much smaller than ℓ, the state-space dimension. Once b is determined it may be substituted back into (13.12) to obtain the solution to the minimization problem.

To dig a little deeper into this calculation we define

$$e^{(m)} = \widehat{v}_{j+1}^{(m)} - \widehat{m}_{j+1}$$

and note that then

$$\widehat{C}_{j+1} = \frac{1}{N}\sum_{m=1}^{N} e^{(m)} \otimes e^{(m)}.$$

Since

$$\widehat{C}_{j+1}a = \frac{1}{N}\sum_{m=1}^{N} b_m e^{(m)}$$

we deduce that this is solved by taking

$$b_m = \langle e^{(m)}, a\rangle.$$

Now note that

$$\frac{1}{2} \mid v - \widehat{v}_{j+1}^{(n)} \mid^2_{\widehat{C}_{j+1}} = \frac{1}{2}\langle a, \widehat{C}_{j+1}a\rangle = \frac{1}{2N}\sum_{m=1}^{N} b_m^2.$$

We define

$$\mathsf{I}_n(b) := \frac{1}{2} \mid y_{j+1}^{(n)} - H\widehat{v}_{j+1}^{(n)} - \frac{1}{N}\sum_{m=1}^{N} b_m H(\widehat{v}_{j+1}^{(m)} - \widehat{m}_{j+1}) \mid^2_{\Gamma} + \frac{1}{2N}\sum_{m=1}^{N} b_m^2. \tag{13.13}$$

We have shown:

Theorem 13.1 (Implementation of EnKF in N Dimensional Subspace)**.** *Given the prediction* ⓟ *defined by (13.8a), the Kalman update formulae* ⓐ *may be found by minimizing $I_n(b)$ with respect to b and substituting into* (13.12)*.*

13.6 Filtering Overview

Having introduced a range of filtering methods in this and the preceding four chapters, it is helpful to summarize what these methods achieve, in a comparitive fashion.

13.6.1 Dynamical Model

The stochastic dynamics model is defined by:

$$v_{j+1} = \Psi(v_j) + \xi_j, \quad j \in \mathbb{R}^+$$

where $\xi_j \sim N(0, \Sigma)$ are independent and identically distributed random variables, also independent of the initial condition $v_0 \sim N(m_0, C_0)$.

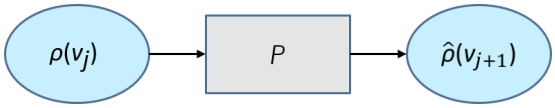


Figure 13 Prediction step.

The pdf of v_j is denoted by $\rho_j(v_j) = \mathbb{P}(v_j)$. This is propagated by the stochastic dynamics model according to the formula $\rho_{j+1} = P\rho_j$ where P is defined in (11.1). This is shown schematically in Figure 13. Note that P is independent of step j because the Markov chain defined by the stochastic dynamics model is time-homogeneous. In the absence of data the probability distribution simply evolves through repeated application of P.

13.6.2 Data Model

The data model is given by

$$y_{j+1} = h(v_{j+1}) + \eta_{j+1},$$

where $\eta_{j+1} \sim N(0, \Gamma)$ are independent and identically distributed random variables, independent of both v_0 and the i.i.d. sequence $\{\xi_j\}$. Given data, we now view the prediction from the dynamical model as a prior which we condition on the data. We write this prior at time $j+1$ as $\widehat{\rho}_{j+1}$, and prediction using the stochastic dynamics model gives $\widehat{\rho}_{j+1} = P\rho_j$.

Using the data to update this prior by Bayes theorem gives an improved estimate of the distribution $\rho_{j+1}(v_{j+1})$ via the formula

$$\rho_{j+1} = L_j \widehat{\rho}_{j+1}$$

where L_j is given by (11.2). This is shown schematically in Figure 14. Note that L_j is nonlinear because $\widehat{\rho}_{j+1}$ appears in both the numerator and the denominator. It depends on j because the data y_{j+1} appears in the equation and this will change with each set of measurements.

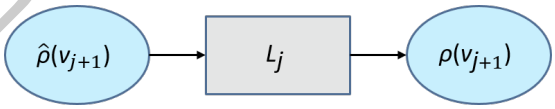


Figure 14 Update Step.

13.6.3 Interaction of the Dynamics and Data

Recall the notation $Y_j = \{y_1, \ldots, y_j\}$ for the collection of all data until time $j+1$. The dynamics and data models described in the two previous subsections can be combined as follows. Using the dynamics, the probability distribution is propagated forward in time

by the prediction step. The data is then used to update the probability distribution at that time in the analysis step. The two can be combined to obtain

$$\rho_{j+1} = L_j P \rho_j$$

The combination of the two steps is shown schematically in Figure 15.

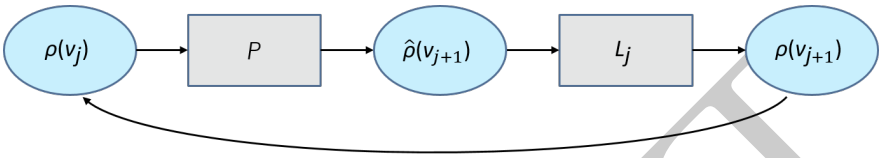


Figure 15 Prediction and update step combined.

13.6.4 Summary of Applicability of Discrete Filtering Methods

There are several filtering methods for performing the prediction and analysis steps. Some methods, such as the bootstrap particle filter, can be applied generally to nonlinear problems. However, others require a linear model ($\Psi(\cdot) = M\cdot$) and/or linear observations ($h(\cdot) = H\cdot$). Some of the methods we have described provably approximate the probability distribution updates. Some just estimate the state and simply use covariance information to weight the relative importance of the predictions from the model and of the data.

The applicability of the methods introduced is summarized in the table below, with respect to linearity/nonlinearity of the dynamics and the observation model. Furthermore **P** is used to denote methods which provably approximate the filtering distributions ρ_j in the large particle limit; **S** denotes methods which only attempt to estimate the state, using the data.

Kalman Filter $\Psi(\cdot) = M\cdot, h(\cdot) = H\cdot$. **P**.

3DVAR general $\Psi, h(\cdot) = H\cdot$. **S**.

Bootstrap Particle Filter general Ψ, h. **P**.

Optimal Particle Filter general $\Psi, h(\cdot) = H\cdot$. **P**.

Extended Kalman Filter general $\Psi, h(\cdot) = H\cdot$. **S**.

Ensemble Kalman Filter general $\Psi, h(\cdot) = H\cdot$. **S**.

Some of these constraints can be relaxed on the setting in which they apply can be relaxed, but the list above describes the methods as we present them in these notes. Furthermore the last two methods can accuratly predict probability distributions in situations where approximate Gaussianty holds; this may be induced by small noise and by large data. However it is importat to appreciate that the raison d'etre of the EnKF is to facilitate the solution of problems with high dimensional state space; typically small

ensemble sizes are used and the algorithm is employed far from the regime in which it is able to provably approximate the distributions, even if they are close to Gaussian. It is for this reason that we prefer to think of the EnKF as an ensemble state estimator.

13.7 Discussion and Bibliography

The development and theory of the extended Kalman filter is documented in the text [55]. A methodology for analyzing evolving probability distributions with small variance, and establishing the validity of the Gaussian approximation, is described in [97]. The use of the ExKF for weather forecasting was proposed in [40]. However the dimension of the state space in most geophysical applications renders the extended Kalman filter impractical. The ensemble Kalman filter provided an innovation with far reaching consequences in geophysical applications, because it allowed for the use of partial empirical correlation information, without the computation of the full covariance. An overview of ensemble Kalman methods may be found in the book [35], including a historical perspective on the subject, originating from papers of Evensen and Van Leeuwen in the mid 1990s [33, 34]; a similar idea was also developed by Houtkamer within the Canadian meteorological service, around the same time; [49, 50].

The presentation of the ensemble Kalman filter as a smart optimization tool is also developed in [70], but the derivation of the update equations in a space whose dimension is that of the ensemble is not described there. The analysis of ensemble methods is difficult and theory is only just starting to emerge. In the linear case the method converges in the large ensemble limit to the Kalman filter [67], but in the nonlinear case the limit does not reproduce the filtering distribution [31]. In any case the primary advantage of ensemble methods is that they can provide good state estimation when the number of particles is *not* large; this subject is discussed in [43, 63, 105, 106].

14 Kalman Smoother

In this chapter, we discuss the Kalman smoother, which refers to the smoothing problem in the case when the dynamics/observation model is linear and subject to additive Gaussian noise, and the initial state distribution is also Gaussian. As with the Kalman filter, it is possible to solve the problem explicitly in this setting, because the posterior is itself a Gaussian. The explicit formulae computed help to build intuition about the smoothing distribution more generally. The mean of the Kalman smoother links directly to the 4DVAR method introduced in chapter 10.

14.1 The Setting

The setting we wish to consider is as follows:

$$\begin{aligned} v_{j+1} &= Mv_j + \xi_j & \xi_j &\sim N(0,\Sigma) \text{ i.i.d.}, \\ y_{j+1} &= Hv_{j+1} + \eta_j & \eta_j &\sim N(0,\Gamma) \text{ i.i.d.}, \end{aligned}$$

with $v_0 \sim N(m_0, C_0)$ independent of the independent i.i.d. sequences $\{\xi_j\}$ and $\{\eta_j\}$. Recall that $Y_j = \{y_1, \dots, y_j\}$ and that the smoothing problem is to determine $\mathbb{P}(v_j|Y_J)$ for $j = 0, \dots, J$. Recall also that the filtering problem is to sequentially update $\mathbb{P}(v_j|Y_j)$ as $j \mapsto j+1$; formulae for this are given in the Kalman filter chapter. The Kalman filter determines the marginal of the Kalman smoother on the coordinate $j = J$, but does not determine the Kalman smoother in its entirety. Thus we derive the formulae for the mean and covariance of the Kalman smoother explicitly.

14.2 Defining Linear System

Let $v = (v_0, \dots, v_J)$ and $y = (y_1, \dots, y_J)$. Using Bayes theorem and the fact that $\{\xi_j\}$, $\{\eta_j\}$ are mutually independent i.i.d. sequences, independent of v_0, we have

$$\mathbb{P}(v|y) \propto \mathbb{P}(y|v)\mathbb{P}(v) = \prod_{j=1}^{J} \mathbb{P}(y_j|v_j) \times \prod_{j=0}^{J-1} \mathbb{P}(v_{j+1}|v_j) \times \mathbb{P}(v_0).$$

Noting that

$$v_{j+1}|v_j \sim N(Mv_j, \Sigma), \quad y_j|v_j \sim N(Hv_j, \Gamma)$$

the Kalman smoothing distribution can be expressed as

$$\mathbb{P}(v|y) \propto \exp(-I(v; y; m_0)), \tag{14.1}$$

where

$$\mathsf{I}(v; y; m_0) = \frac{1}{2}|v_0 - m_0|^2_{C_0} + \frac{1}{2}\sum_{j=0}^{J-1} |v_{j+1} - Mv_j|^2_{\Sigma} + \frac{1}{2}\sum_{j=0}^{J-1} |y_{j+1} - Hv_{j+1}|^2_{\Gamma}. \tag{14.2}$$

Theorem 14.1 (Characterization of the Kalman Smoother). *Assume $\Sigma, C_0, \Gamma > 0$. Then $\mathbb{P}(v|y)$ is Gaussian distributed with a tridiagonal precision matrix $L > 0$ solving $Lm = r$, mean*

m, defined as follows:

$$L = \begin{bmatrix} L_{0,0} & L_{0,1} & & & & \\ L_{1,0} & L_{1,1} & \dots & & 0 & \\ 0 & \dots & \dots & & & \\ & 0 & \dots & \dots & \dots & \\ & & & \dots & L_{J-1,J-1} & L_{J-1,J} \\ & & & & L_{J,J-1} & L_{J,J} \end{bmatrix} \tag{14.3}$$

with

$$\begin{aligned} L_{0,0} =& C_0^{-1} + M^T\Sigma^{-1}M, \\ L_{j,j} =& \Sigma^{-1} + M^T\Sigma^{-1}M + H^T\Gamma^{-1}H, \quad 1 \le j \le J-1, \\ L_{J,J} =& \Sigma^{-1} + H^T\Gamma^{-1}H, \\ L_{j,j+1} =& -M^T\Sigma^{-1}, \quad 1 \le j \le J-1, \\ r_0 =& C_0^{-1}m_0, \\ r_j =& H^T\Gamma^{-1}y_j, \quad 1 \le j \le J. \end{aligned}$$

Proof. We may write $\mathsf{I}(v, y, m_0) = \frac{1}{2}|L^{1/2}(v-m)|^2 + q$ with q independent of v, by definition. To find L, we will derive the Hessian of $I(v, y, m_0)$ w.r.t. v; L is the Hessian of I. Note that this tells us that

$$\begin{aligned} L_{0,0} =& \partial^2_{v_0}\mathsf{I}(v, y, m_0) = C_0^{-1} + M^T\Sigma^{-1}M, \\ L_{j,j} =& \partial^2_{v_j}\mathsf{I}(v, y, m_0) = \Sigma^{-1} + M^T\Sigma^{-1}M + H^T\Gamma^{-1}H, \\ L_{J,J} =& \partial^2_{v_J}\mathsf{I}(v, y, m_0) = \Gamma^{-1} + H^T\Gamma^{-1}H, \\ L_{j-1,j} =& L_{j,j-1} = \partial^2_{v_j,v_{j-1}}\mathsf{I}(v, y, m_0) = -M^T\Sigma^{-1}. \end{aligned}$$

Otherwise, for all other values of indices $\{k, l\}$, $L_{k,l} = 0$. This proves that the matrix L has a tridiagonal form.

Now we focus on finding m. We have that $\nabla_v\mathsf{I}(v, y, m_0) = L(v-m)$, so that $-\nabla_v\mathsf{I}(v, y, m_0)|_{v=0} = Lm$. Thus, we find r as,

$$\begin{aligned} r_0 =& -\nabla_{v_0}\mathsf{I}(v, y, m_0)|_{v=0} = -(-C_0^{-1}m_0) = C_0^{-1}m_0, \\ r_j =& -\nabla_{v_j}\mathsf{I}(v, y, m_0)|_{v=0} = -(-H^T\Gamma^{-1}y_j) = H^T\Gamma^{-1}y_j. \end{aligned}$$

We have shown that L is symmetric and that $L \ge 0$; to prove that L is a precision matrix, we need to show that $L > 0$. Suppose we take $y = 0$ and $m_0 = 0$, so that every term in the expansion of $\mathsf{I}(v; 0; 0)$ involves v. It is evident that $\mathsf{I}(v; 0; 0) = v'Lv$. Suppose that $v'Lv = 0$ for some nonzero v. Then by positive-definiteness of C_0, Σ, and Γ, it must be that $v_0 = 0$ and $v_{j+1} = Mv_j$ for $j = 0, 1, , \dots, J$. Equivalently, we must have $v = 0$. This proves that L must be positive-definite. □

Remark 14.2. We note that the results of solving the smoothing problem in this case may be reformulated as solving an equivalent optimization problem. Finding the posterior mean of $\mathbb{P}(v|y)$ as the unique minimizer of $\mathsf{I}(v, y, m_0)$. This is equivalent to finding the MAP estimator. □

14.3 Solution of the Linear System

We may also obtain m via Gaussian elimination, using the block tridiagonal structure of L, as follows. First we form the matrix sequence $\{L_j\}$:

$$\begin{aligned} L_0 =& L_{0,0} \\ L_{j+1} =& L_{j+1,j+1} - \Sigma^{-1} M^T L_j^{-1} M^T \Sigma^{-1}, \quad j = 0, \ldots, J-1; \end{aligned} \tag{14.4}$$

and we form the vector sequence $\{z_j\}$:

$$\begin{aligned} z_0 =& C_0^{-1} m_0, \\ z_{j+1} =& H^T \Gamma^{-1} y_{j+1} - \Sigma^{-1} M^T L_j^{-1} z_j. \end{aligned}$$

We may then read off m_J from the equation $L_J m_J = Z_J$ and finally we perform back-substitution to obtain

$$L_j m_j = z_j - L_{j,j+1} m_{j+1}, \quad j = J-1, .., 1.$$

Note that m_J found this way coincides with the mean of the Kalman filter at $j = J$.

Proposition 14.3. *The matrices $\{L_j\}$ in (14.4) are positive definite*

Proof. The proof of this theorem relies on the following two lemmas:

Lemma 14.4. *If*

$$X := \begin{bmatrix} X_1 & \times & \times & \times \\ \times & X_2 & \times & \times \\ \times & \times & ... & \times \\ \times & \times & \times & X_n \end{bmatrix}$$

is positive-definite symmetric then X_i is positive-definite symmetric for all $i \in \{1, \ldots, n\}$.

Lemma 14.5. *Let P be a block upper(lower) triangular matrix with identity on the diagonal. Then, P is an invertible matrix.*

Using Lemma 14.4, we deduce that $L_0 = L_{0,0}$ is positive-definite symmetric. Consider the matrix $P \in \mathbb{R}^{N(J+1) \times N(J+1)}$ defined as :

$$P = \begin{bmatrix} I & 0 & & 0 \\ -L_{1,0} L_0^{(-1)} & I & ... & ... \\ 0 & & ... & 0 \\ & & 0 & I \end{bmatrix}.$$

We compute

$$PLP^T = \begin{bmatrix} L_0 & 0 & & & 0 \\ 0 & L_1 & L_{1,2} & \dots & 0 \\ & L_{2,1} & L_{2,2} & \dots & \\ & & & & \\ 0 & & & L_{J-1,J-1} & L_{J-1,J} \\ 0 & & & L_{J,J-1} & L_{J,J} \end{bmatrix}.$$

By using the Lemma 14.4,

$$\tilde{L} = \begin{bmatrix} L_1 & L_{1,2} & \dots & 0 \\ L_{2,1} & L_{2,2} & \dots & \\ & & & \\ & & L_{J-1,J-1} & L_{J-1,J} \\ & & L_{J,J-1} & L_{J,J} \end{bmatrix}$$

is positive definite, and so is L_0.

Lemma 14.4 gives us that L_1 positive definite as follows. By Lemma 14.5 the following matrix, which we will use to transform the above matrix, is invertible.

$$P_2 = \begin{bmatrix} I & 0 & 0 \\ -L_{2,1}L_1^{-1} & I & \dots & \dots \\ & & \dots & 0 \\ 0 & & & I \end{bmatrix}.$$

Thus we have

$$P_2 \tilde{L} P_2^T = \begin{bmatrix} L_1 & 0 & & & 0 \\ 0 & L_2 & L_{2,3} & \dots & \\ & L_{3,2} & L_{3,3} & & 0 \\ & & & & \\ & & & L_{J-1,J-1} & L_{J-1,J} \\ 0 & & & L_{J,J-1} & L_{J,J} \end{bmatrix},$$

giving the positive definiteness of L_1. We may iterate our reasoning to argue that all the L_j are similarly positive definite. □

14.4 Discussion and Bibliography

The subject of Kalman smoothing is overviewed in the text [5]; see also [92]. A link between the standard implementation of the smoother and Gauss-Newton methods for MAP estimation is made in [8]. Fur further details on the Kalman smoother, in both discrete and continuous time, see [70] and [44].

15 Filtering Approach to the Inverse Problem

In this final chapter we demonstrate how the two separate themes that underpin this course, inverse problems and data assimilation, may be linked. This opens up the possibility of transfering ideas from filtering into the setting of quite general inverse problems. In the first section we describe the general, abstract, connection and introduce the revolutionary idea of sequential Monte Carlo methods (SMC). In the second section we analyze the concrete case of applying the EnKF to solve an inverse problem, leading to ensemble Kalman inversion (EKI). In the final section we link the EKI methodology to SMC.

15.1 General Formulation

Recall the inverse problem of finding $u \in \mathcal{X}$ from $y \in \mathbb{R}^J$ where

$$y = G(u) + \eta,\ \eta \sim N(0, \Gamma_0) \tag{15.1}$$

and the related loss function

$$\varphi_0(u) = \frac{1}{2}|y - G(u)|^2_{\Gamma_0}.$$

The reason for writing Γ_0, rather than Γ, will become apparent below and will also be exploited in the third section of this chapter where we study EKI. The reason for considering $\mathcal{X}$ rather than $\mathbb{R}^N$ as in the earlier chapters on inversion is twofold: (i) we avoid notational conflict with N the number of ensemble members; (ii) we emphasize that the ideas of this chapter apply quite generally when unknown u lies in in a Hilbert space.

If we put a prior ρ_0 on the unknown u then the posterior takes the form

$$\rho(u) = \frac{1}{Z}\exp(-\varphi_0(u))\rho_0(u).$$

Let $J \in \mathbb{N}$ and choose h so that $Jh = 1$. Then define the family of pdfs $\{\rho_j\}_{j=0}^J$ by

$$\rho_j(u) = \frac{1}{Z_j}\exp(-jh\varphi_0(u))\rho_0(u).$$

It follows that $\rho_J = \rho$ and, furthermore, we may update the sequence $\{\rho_j\}_{j=0}^J$ sequentially using the formula

$$\rho_{j+1}(u) = \frac{Z_j}{Z_{j+1}}\exp(-h\varphi_0(u))\rho_j(u).$$

This simply corresponds to application of Bayes' Theorem to the inverse problem

$$y = G(u) + \eta,\ \eta \sim N(0, \Gamma), \tag{15.2}$$

with $\Gamma = \frac{1}{h}\Gamma_0$ and with prior ρ_j. With this choice of Γ the identity (15.2) gives the likelihood proportional to $\exp(-h\varphi_0(u))$. The update may be written

$$\rho_{j+1}(u) = L\rho_j,$$

noting that the likelihood operator L is nonlinear, corresponding to multiplication by $\exp(-h\varphi_0(u))$ and then normalization to a pdf.

If we let P_j denote any Markov kernel for which ρ_j is invariant then we obtain

$$\rho_{j+1} = LP_j\rho_j.$$

This update formula should be compared with the filtering update formula

$$\rho_{j+1} = L_jP\rho_j$$

introduced and used in earlier chapters.

Whilst the $j-$dependence in Bayes rule and the Markov kernel is interchanged between the inverse problem and the filtering problem, this makes little material difference to implementation. Firstly, once a Markov kernel P is identified under which $\rho = \rho_J$ is invariant, it can be easily adapted to find a family of kernels P_j under which ρ_j is invariant, simply by rescaling the observation covariance. Secondly the fact that the data y is fixed, rather than changing at each step, makes little difference to the implementation. This perspective opens up an entire field, known as *sequential Monte Carlo*, in which ideas from filtering may be transferred to other quite different problems, including Bayesian inversion, as explained here.

Note however (see the discussion at the end of section 13.6) that of the filtering methods we have introduced so far only the bootstrap filter applies directly to the case of *nonlinear* observation operators; since G is in general nonlinear this means that extra ideas are required to implement the optimal particle filter, 3DVAR, ExKF and EnKF. One approach is to replace the sequential optimization principles by the non-quadratic optimization problem required for nonlinear G; we do not discuss this idea in any detail but it is a viable option. Another approach is to use the linearization technique which we now outline in the context of EKI.

15.2 Ensemble Kalman Inversion

We now make a detour into the subject of how to use filters to estimate parameters u from data y satisfying (15.2). The approach we study, and which we will relate to sequential Monte Carlo in the next section, is to introduce an artificial time dynamic. This can be done quite simply as follows: we write

$$\begin{aligned} u_{j+1} &= u_j, \\ y_{j+1} &= G(u_{j+1}) + \eta_{j+1}, \end{aligned}$$

and we can think of finding the filtering distribution on $u_j|Y_j$. We discuss how to relate the y_j to the one data point y below. For now we take $\eta_j \sim N(0,\Gamma)$ but we will revisit this choice in the next section, linking the problem to solution of (15.1). Because the observation operator G is, in general, nonlinear, this does not render our system in a form where we can readily apply the EnKF. To this end we introduce a new variable w_j and rewrite the filter as:

$$\begin{aligned} u_{j+1} &= u_j, \\ w_{j+1} &= G(u_j), \\ y_{j+1} &= w_{j+1} + \eta_{j+1}. \end{aligned}$$

We introduce the new variables $v = (u, w)^T$, nonlinear map $\Psi(v) = (u, G(u))^T$ and linear operators $H = [0, I], H^\perp = [I, 0]$. Then if we write $v_j = (u_j, w_j)^T$ we may write the dynamical system in the form

$$v_{j+1} = \Psi(v_j) \tag{15.3a}$$
$$y_{j+1} = Hv_{j+1} + \eta_{j+1}. \tag{15.3b}$$

We note that

$$Hv = w, \quad H^\perp v = u.$$

Remark 15.1. Typically either of the following are used to construct artificial data $\{y_j\}$ for the filtering algorithm, given a single instance of data y:

$$y_{j+1} = \begin{cases} y \text{ (unperturbed obervations)} \\ y + \bar{\eta}_{j+1}, \bar{\eta}_{j+1} \sim N(0, \Gamma) \text{ (perturbed observations)} \end{cases}$$

The first choice is natural if viewing the algorithm as a sequential optimizer; the latter, in the linear case, is natural when seeking to draw samples from the posterior. □

We now apply the EnKF to the dynamics/data model (15.3). We obtain, for $n = 1, \ldots, N$,

$$\widehat{v}_{j+1}^{(n)} = \Psi(v_j^{(n)}), \tag{15.4}$$
$$\bar{v}_{j+1} = \frac{1}{N} \sum_{n=1}^{N} \widehat{v}_{j+1}^{(n)}, \tag{15.5}$$
$$\widehat{C}_{j+1} = \frac{1}{N} \sum_{n=1}^{N} (\widehat{v}_{j+1}^{(n)} - \bar{v}_{j+1}) \otimes (\widehat{v}_{j+1}^{(n)} - \bar{v}_{j+1}), \tag{15.6}$$
$$v_{j+1}^{(n)} = (I - K_{j+1}H)\widehat{v}_{j+1}^{(n)} + K_{j+1} y_{j+1}^{(n)} \tag{15.7}$$

with the Kalman gain

$$K_{j+1} = \widehat{C}_{j+1} H^T S_{j+1},$$
$$S_{j+1} = (HC_{j+1}H^T + \Gamma)^{-1}$$

Now we may simplify these expressions by using the specific Ψ, v, H arising in the inverse problem:

$$\widehat{C}_{j+1} = \begin{bmatrix} C_{j+1}^{uu} & C_{j+1}^{uw} \\ (C_{j+1}^{uw})^T & C_{j+1}^{ww} \end{bmatrix} \quad \bar{v}_{j+1} = \begin{pmatrix} \bar{u}_{j+1} \\ \bar{w}_{j+1} \end{pmatrix}$$

Here

$$\bar{u}_{j+1} = \frac{1}{N} \sum_{n=1}^{N} u_j^{(n)}, \ \bar{w}_{j+1} = \frac{1}{N} \sum_{n=1}^{N} G(u_j^{(n)}) := \bar{G}_j$$

and

$$C_{j+1}^{uw} = \frac{1}{N}\sum_{n=1}^{N}(u_j^{(n)} - \bar{u}_{j+1}) \otimes (G(u_j^{(n)}) - \bar{G}_j),\ C_{j+1}^{ww} = \frac{1}{N}\sum_{n=1}^{N}(G(u_j^{(n)}) - \bar{G}_j) \otimes (G(u_j^{(n)}) - \bar{G}_j).$$

There is a similar expression for C_{j+1}^{uu}, but as we will show in what follows it is not needed for the unknown parameter u update formula. Noting that, because of the structure of H, $S_{j+1} = (C_{j+1}^{ww} + \Gamma)^{-1}$ we obtain

$$K_{j+1} = \begin{pmatrix} C_{j+1}^{uw}(C_{j+1}^{ww} + \Gamma)^{-1} \\ (C_{j+1}^{ww})(C_{j+1}^{ww} + \Gamma)^{-1} \end{pmatrix} \tag{15.8}$$

Combining equation (15.8) with the update equation within (15.4) it follows that

$$\{v_j^{(n)}\}_{n\in\{1,N\}} \to \{v_{j+1}^{(n)}\}_{n\in\{1,N\}}$$

and

$$\{H^{\perp}v_j^{(n)}\}_{n\in\{1,N\}} \to \{H^{\perp}v_{j+1}^{(n)}\}_{n\in\{1,N\}}$$

and hence that

$$u_{j+1}^{(n)} = H^{\perp}v_{j+1}^{(n)} = u_j^{(n)} + C_{j+1}^{uw}(C_{j+1}^{ww} + \Gamma)^{-1}(y_{j+1}^{(n)} - G(u_j^{(n)})).$$

Thus we have derived the EKI step

$$u_{j+1}^{(n)} = u_j^{(n)} + C_{j+1}^{uw}(C_{j+1}^{ww} + \Gamma)^{-1}(y_{j+1}^{(n)} - G(u_j^{(n)})). \tag{15.9}$$

The full algorithm is described below

Algorithm 15.1 Algorithm for Ensemble Kalman Inversion

1: **Input**: Initial distribution $\mathbb{P}(u_0) = \rho_0$, observations Y_J, number of particles N
2: **Initial Sampling**: Draw N particles $u_0^{(n)} \sim \rho_0$ so that $\rho_0^N = S^N \rho_0$
3: **Subsequent Sampling** For $j = 0, 1, \ldots, J-1$, perform

1. Set $\bar{u}_{j+1} = \frac{1}{N} \sum_{n=1}^N u_j^{(n)}$
2. Set $\bar{G}_j = \frac{1}{N} \sum_{n=1}^N G(u_j^{(n)})$
3. Set $C_{j+1}^{uw} = \frac{1}{N} \sum_{n=1}^N (u_j^{(n)} - \bar{u}_{j+1}) \otimes (G(u_j^{(n)}) - \bar{G}_j)$
4. Set $C_{j+1}^{uw} = \frac{1}{N} \sum_{n=1}^N (u_j^{(n)} - \bar{u}_{j+1}) \otimes (G(u_j^{(n)}) - \bar{G}_j)$, $C_{j+1}^{ww} = \frac{1}{N} \sum_{n=1}^N (G(u_j^{(n)}) - \bar{G}_j) \otimes (G(u_j^{(n)}) - \bar{G}_j)$
5. $u_{j+1}^{(n)} = u_j^{(n)} + C_{j+1}^{uw} (C_{j+1}^{ww} + \Gamma)^{-1} (y_{j+1}^{(n)} - G(u_j^{(n)}))$.

4: **Output**: N particles $u_J^1, u_J^2, \ldots, u_J^N$

Remark 15.2. This algorithm may be viewed as a derivative-free optimization method, within the broad class that contains genetic, swarm or ant colony optimization. □

The algorithm has the following invariant subspace property.

Theorem 15.3. *(Space of Ensemble) Define $\mathcal{A} = span(u_0^{(n)})_{n \in \{1, \cdots, N\}}$, then for all j in $\{0, \ldots, J\}$ and for all n in $\{1, \cdots, N\}$, $u_j^{(n)}$ defined by the iteration (15.9) lie in $\mathcal{A}$.*

Proof. The proof proceeds by induction. We let $\mathfrak{u}_j \in \mathcal{X}^N$ denote the collection of $\{u_j^{(n)}\}_{n \in \{1, \ldots, N\}}$. At first, let j be in $\{0, \ldots, J\}$, n and m in $\{1, \ldots, N\}$ and define the following two quantities:

$$d_j^{(n)} := (C_{j+1}^{ww} + \Gamma)^{-1} (y_{j+1}^{(n)} - G(u_j^{(n)}))$$
$$D_{mn}(\mathfrak{u}_j) := -\langle d_j^{(n)}, G(u_j^{(m)}) - \bar{G}_j \rangle$$

Notice that the $d_j^{(n)}$ are elements of the data space $\mathbb{R}^J$, whereas the $D_{mn}(\mathfrak{u}_j)$ are scalar quantities. Furthermore notice that

$$u_{j+1}^{(n)} = u_j^{(n)} - \frac{1}{N} \sum_{m=1}^N D_{mn}(\mathfrak{u}_j)(u_j^{(m)} - \bar{u_j}).$$

From the definition of $\bar{G}_j$ and bilinearity of the inner product it follows that $\sum_{m=1}^N D_{mn}(\mathfrak{u}_j) = 0$. Therefore the update expression may be rewritten as, for all $j \in \{0, \ldots, J\}$ and $n \in \{1, \ldots, N\}$,

$$u_{j+1}^{(n)} = u_j^{(n)} - \frac{1}{N} \sum_{m=1}^N D_{mn}(\mathfrak{u}_j) u_j^{(m)}.$$

Hence if the property holds for all the particles of the time step j, it will clearly be the case for all the particles at time step $j+1$. □

Remark 15.4. The choice of the initial ensemble of particles $\{u_0^{(n)}\}_{n\in\{1,\dots,N\}}$ is thus crucial to the performance of the EnKF, since the algorithm remains in the initial space of the initial ensemble. In the setting of Bayesian inverse problems the initial ensemble is frequently created by drawing from the prior. Alternatively, if the prior is Gaussian, the first N Karhunen-Loeve eigenfunctions may be used, ordered by decreasing variance contribution. More generally any truncated basis for the space $\mathcal{X}$ is a natural initial ensemble. However the question of how to adaptively learn a good choice of ensemble subspace, in response to observed data, is unexplored and potentially fruitful. □

Remark 15.5. We describe an alternative way to approach the derivation of the EKI update formulae. We apply Theorem 13.1 with the specific structure arising from the dynamical system used in EKI. To this end we define

$$\mathsf{I}_n(b) := \frac{1}{2} \mid y_{j+1}^{(n)} - G(u_j^{(n)}) - \frac{1}{N}\sum_{m=1}^{N} b_m (G(u_j^{(m)}) - \bar{G}_j) \mid_{\Gamma}^2 + \frac{1}{2N}\sum_{m=1}^{N} b_m^2. \tag{15.10}$$

Once this quadratic form has been minimized with respect to b then the upate formula (13.12) gives

$$u_{j+1}^{(n)} = u_j^{(n)} + \frac{1}{N}\sum_{m=1}^{N} b_m (u_j^{(m)} - \bar{u}_j),$$
$$w_{j+1}^{(n)} = G(u_j^{(n)}) + \frac{1}{N}\sum_{m=1}^{N} b_m (G(u_j^{(m)}) - \bar{G}_j),$$

(Note that the vector $\{b_m\}$ depends on n; we have supressed this dependence for notational convenience). Theorem 15.3 is an immediate consequence of this structure. □

15.3 Linking Ensemble Kalman Inversion and SMC

He we link the two preceding sections. In the first subsection we describe how ensemble Kalman inversion may be linked with SMC through a particular scaling of Γ with h. We then take the limit $h \to 0$ and obtain a system or ordinary differential equations. In the second subsection we study the resulting algorithm in the case in which G is linear, leading to insight into the EKI algorithm in the iterated context.

15.3.1 Continuous Time Limit

Although we have emphasized the optimization perspective on ensemble methods, we may think of one-step of ensemble Kalman inversion as approximating the filtering mapping $\rho_j = \mathbb{P}(u_j|Y_j) \mapsto \rho_{j+1} = \mathbb{P}(u_{j+1}|Y_{j+1})$; this will be a good approximation with measures are close to Gaussian and when the number of ensemble members N is

large. In order to link this to the mapping $\rho_{j+1} = L\rho_j$ in the presentation of SMC in the first section of this chapter, we set $\Gamma = \frac{1}{h}\Gamma_0$. Let $\mathfrak{u} \in \mathcal{X}^N$ denote the collection of $\{u^{(n)}\}_{n\in\{1,\dots,N\}}$. Now define

$$D^0_{mn}(\mathfrak{u}) := \langle G(u^{(n)}) - y, G(u^{(m)}) - \bar{G}\rangle_{\Gamma_0},$$

where

$$\bar{G} := \frac{1}{N}\sum_{m=1}^{N} G(u^{(m)}).$$

Note that then, to leading order in $h \ll 1$,

$$D_{mn}(\mathfrak{u}) \approx h\langle G(u^{(n)}) - y, G(u^{(m)}) - \bar{G}\rangle_{\Gamma_0} = hD^0_{mn}(u).$$

It follows that, also to leading order in h,

$$u^{(n)}_{j+1} \approx u^{(n)}_j - \frac{h}{N}\sum_{m=1}^{N} D^0_{mn}(\mathfrak{u}_j)u^{(m)}_j$$

where $\mathfrak{u}_j$ is as defined analogously to the previous section to denote the collection of particles at discrete time j. Then letting $h \longrightarrow 0$ yields:

$$\frac{du^{(n)}}{dt} = -\frac{1}{N}\sum_{m=1}^{N} D^0_{mn}(\mathfrak{u})u^{(m)} \tag{15.12}$$

Remark 15.6. Notice that equation (15.12) has families of fixed points where: (i) either the particles fit the data exactly (which corresponds to the left hand side in the inner product defining $D^0_{mn}(u)$ being zero); or (ii) all particles collapse on their mean value (which corresponds to the right hand side of the same inner product). This suggests that the system of ordinary differential equations which describes the behaviour of ensemble Kalman inversion is driven by two desirable attributes: matching the data and achieving consensus. □

15.3.2 Linear Setting

Now suppose $G(\cdot)$ is a linear map denoted $A\cdot$, and define

$$\bar{u} = \frac{1}{N}\sum_{m=1}^{N} u^{(m)}.$$

In this setting we have

$$\begin{aligned}\frac{du^{(n)}}{dt} &= -\frac{1}{N}\sum_{m=1}^{N}\langle Au^{(n)} - y, A(u^{(m)} - \bar{u})\rangle_{\Gamma_0} u^{(m)} \\ &= -\frac{1}{N}\sum_{m=1}^{N}\langle Au^{(n)} - y, A(u^{(m)} - \bar{u})\rangle_{\Gamma_0} (u^{(m)} - \bar{u}) \\ &= -\frac{1}{N}\sum_{m=1}^{N}\langle A^*\Gamma_0^{-1}(Au^{(n)} - y), u^{(m)} - \bar{u}\rangle (u^{(m)} - \bar{u}) \\ &= -\frac{1}{N}\sum_{m=1}^{N}(u^{(m)} - \bar{u}) \otimes (u^{(m)} - \bar{u})(A^*\Gamma_0^{-1}(Au^{(n)} - y)).\end{aligned}$$

If we define

$$C(\mathfrak{u}) = \frac{1}{N}\sum_{m=1}^{N}(u^{(m)} - \bar{u}) \otimes (u^{(m)} - \bar{u}) \tag{15.13}$$

then we find that, for $n = 1, \ldots, N$,

$$\begin{aligned}\frac{du^{(n)}}{dt} &= -C(\mathfrak{u})(A^*\Gamma_0^{-1}(Au^{(n)} - y) \\ &= -C(\mathfrak{u})\nabla\Phi(u^{(n)}),\end{aligned}$$

where

$$\Phi(u) = \frac{1}{2}|y - Au|^2_{\Gamma_0}.$$

This corresponds to a gradient descent in the subspace defined by the initial ensemble. In particular

$$\begin{aligned}\frac{d}{dt}\Phi(u^{(n)}(t)) &= \langle \nabla\Phi(u^{(n)}(t)), \frac{du^{(n)}}{dt}(t)\rangle \\ &= -|C(u)^{1/2}\nabla\Phi(u^{(n)}(t))|^2 \leq 0\end{aligned}$$

demonstrating that the loss function Φ is decreasing along each trajectory associated to each ensemble member.

We note also that, if we define $e^{(n)} = u^{(n)} - \bar{u}$, then

$$\frac{de^{(n)}}{dt} = -C(\mathfrak{u})A^*\Gamma_0^{-1}Ae^{(n)}.$$

Because

$$C := C(\mathfrak{u}) = \frac{1}{N}\sum_{n=1}^{N} e^{(n)} \otimes e^{(n)}$$

it follows that

$$\frac{dC}{dt} = -2CA^*\Gamma_0^{-1}AC.$$

Note further that if C is invertible (requires N at least as big as the dimension of $\mathcal{X}$) then the inverse P satisfies

$$\frac{dP}{dt} = 2A^*\Gamma_0^{-1}A$$

so that $\|P\| \to 0$ as $t \to \infty$. This fact, suitably interpreted, even in the case where C is only invertible on a subspace, drives the covariance to zero and causes ensemble collapse – consensus – and at the same time Φ is minimized over an appropriate subspace.

15.4 Discussion and Bibliography

The idea of using particle filters to sample general distributions, including those arising in Bayesian inversion, maybe be found in [26]; a recent application to a Bayesian inverse problem, which demonstrated the potential of the methodology in that context, is [60]. A simple proof of convergence of the method may be found in [10]; it is based on the proof described in [93] for the standard bootstrap particle filter.

The use of the ensemble Kalman filter for parameter estimation was introduced in the papers [75, 7] in which a physical dynamical model was appended with trivial dynamics for the parameters in order to estimate them; the idea was extended to learn an entire field of parameter values in [84]. The paper [99] was the first to do what we do here, namely to consider all the data at once and a single mapping of unknowns parameters to data. In that paper only one iteration is used; the papers [19, 29] demonstrated how iteration could be useful. See also the book [87] for the use of ensemble Kalman methods in oil reservoir simulation.

Development of the method for general inverse problems was undertaken in [51], and further development of iterative methods is described in [53, 54]. Analysis of ensemble inversion was undertaken in [98], including the continuous-time limit and gradient flow structure described here. The potential for ensemble inversion in the context of hierarchical Bayesian methods is demonstrated in [18]. The idea that we explain here, of obtaining an evolution equation for the covariance which is satisfied exactly by ensemble methods, appears in the remarkable papers [94, 11], and in the other papers of Bergemann and Reich referenced therein. Their work is at the heart of the analysis in [98] which demonstrates the ensemble collapse and approximation properties of the ensemble method when applied to linear inverse problems.

References

[1] Abarbanel, H. (2013). Predicting The Future: Completing Models Of Observed Complex Systems. *Springer*.

[2] Agapiou, S., M. Burger, M. Dashti, and T. Helin (2018). Sparsity-promoting and edge-preserving maximum a posteriori estimators in non-parametric Bayesian inverse problems. *Inverse Problems*, 34(4).

[3] Agapiou, S., S Larsson, and A.M. Stuart (2013). Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Processes and their Applications*, 123(10):3828–3860.

[4] Agapiou, S., O. Papaspiliopoulos, D. Alonso, A.M. Stuart (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431.

[5] Anderson, B. and J. Moore (1979). Optimal filtering. *Englewood Cliffs*, 21:22–95.

[6] Anderson, C (2014). Monte Carlo methods and importance sampling *Lecture Notes for Statistical Genetics*.

[7] Anderson, J. (2001). An Ensemble Adjustment Kalman Filter for Data Assimilation *Monthly Weather Review*, 129:2284–2903.

[8] Bell, B. The iterated Kalman smoother as a Gauss–ewton method (1994). *SIAM Journal on Optimization*, 4(3):626–636.

[9] Bickel, P, B. Li and T. Bengtsson. Pushing the limits of contemporary statistics: Contributions in honor of Jayanta K. Ghosh: Sharp failure rates for the bootstrap particle filter in high dimensions, *Institute of Mathematical Statistics*, 318–329.

[10] Beskos.A., A. Jasra, K. Law, R. Tempone, and Y. Zhou. Multilevel Sequential Monte Carlo Samplers. *Stochastic Processes and their Applications*, 127(5):1417–1440.

[11] Bergemann, K. and S. Reich. An ensemble Kalman-Bucy filter for continuous data assimilation. *Meteorologische Zeitschrift*, 127(5):1417–1440.

[12] Bissiri, P., C. Holmes, and S. Walker (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130.

[13] Bishop, C (2006). Pattern Recognition and Machine Learning. *SIAM*.

[14] Brett, C., K. Lam, K. Law, D. McCormick, M. Scott, and A.M. Stuart (2013). Accuracy and stability of filters for dissipative PDE's. *Physica D: Nonlinear Phenomena*, 245(1):34–45.

[15] Bröcker, J. (2013). Existence and uniqueness for four-dimensional variational data assimilation in discrete time. *SIAM Journal on Applied Dynamical Systems*, 16(1):361–374.

[16] Brooks, S., A. Gelman, G. Jones and X. Meng (2011). Handbook of Markov chain Monte Carlo. *CRC press*.

[17] Carrassi, A., M. Bocquet, L. Bertino and G. Geir (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 9(5):e535.

[18] Chada, N., M. Iglesias, L. Roininen, and A.M. Stuart (2017). Parameterizations for ensemble Kalman inversion. Inverse Problems, 34 (2018) 055009.

[19] Chen, Y. and D. Oliver (2002). Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Mathematical Geosciences*, 44(1): 1–26.

[20] Crisan, D., P. Del Moral, and T. Lyons (1998). Discrete filtering using branching and interacting particle systems. *Université de Toulouse. Laboratoire de Statistique et Probabilités [LSP]*.

[21] Cotter, S., M. Dashti, and A.M. Stuart (2010). Approximation of Bayesian inverse problems for PDE's. *SIAM Journal on Numerical Analysis*, 48(1):322–345.

[22] Cotter, S., G. Roberts, A.M. Stuart, and D. White (2013). MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446.

[23] Dashti, M, K. Law, A.M. Stuart, and J. Voss (2013). Map estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Problems*, 29(9):095017.

[24] Dashti, M. and A.M. Stuart (2017). The Bayesian approach to inverse problems. *Handbook of Uncertainty Quantification*, 311–428.

[25] Del Moral, P. Feynman-Kac formulae (2004). In *Genealogical and Interacting Particle Systems with Applications*, pages 47–93. Springer.

[26] Del Moral, P., A. Doucet, and A. Jasra (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

[27] Doucet, A., N. Freitas, and N. Gordon (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo Methods in Practice*, 3–14, Springer.

[28] Doucet, A., S. Godsill, and C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208.

[29] Emerick, A. and A. Reynolds (2013). Investigation of the sampling performance of ensemble-based methods with a simple reservoir model. *Computational Geosciences*, 17(2):325–350.

[30] Engl, H., M. Hanke, and A. Neubauer (1996). Regularization of inverse problems. *Springer Science & Business Media*, 375.

[31] Ernst, O., B. Sprungk, and H. Starkloff (2015). Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):823–851.

[32] Evans, M. and T. Swartz (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, 254–272.

[33] Evans, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(c5): 10143–10162.

[34] Evans, G. and P. Van Leeuwen (1996). Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model. *Monthly Weather Review*, 124(1): 85–96.

[35] Evensen, G (2009). *Data Assimilation: The Ensemble Kalman Filter*. Springer Science & Business Media.

[36] Franklin, J (1970). *Well-posed stochastic extensions of ill-posed linear problems*. *Journal of mathematical analysis and applications*, 31(3):682–716.

[37] Fisher, M., J., Nocedal, Y. Trémolet, and S. Wright (2009). Data assimilation in weather forecasting: a case study in PDE-constrained optimization. *Optimization and Engineering*, 10(3):409–426.

[38] Gamerman, D and H. Lopes (2006). Markov chain Monte Carlo: stochastic simulation for Bayesian inference. *CRC Press*.

[39] Gibbs, A. and F. Su (2002). On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435.

[40] Ghil, M., S. Cohn, J. Tavantzis, and K. Bube (1981). Application of estimation theory to numerical weather prediction. *Dynamic Meteorology: Data Assimilation Methods*, New York: Springer.

[41] Gillijns, S., O. Mendoza, J. Chandrasekar, B. Moor, D. Bernstein, and A. Ridley (2006). What is the ensemble Kalman filter and how well does it work? *American Control Conference*.

[42] Giné, E. and R. Nickl (2015). Mathematical foundations of infinite-dimensional statistical models. *Cambridge University Press*,40.

[43] Gottwald, G. and A. Majda (2013). A mechanism for catastrophic filter divergence in data assimilation for sparse observation networks. *Nonlinear Processes in Geophysics*, 20(5):705–712.

[44] Hairer, M. A. Stuart, J. Voss, and P. Wiberg (2005). Analysis of SPDEs arising in path sampling. Part I: The Gaussian case. *Communications in Mathematical Sciences*, 3(4):587–603.

[45] Harvey, A. (1990). Forecasting, structural time series models and the Kalman filter. *Cambridge university press*.

[46] Hammersley, J. and D. Handscomb (1964). Percolation processes. *Monte Carlo Methods*, 134–141. Springer.

[47] Hayden, K., E. Olson and E. Titi (2011). Discrete data assimilation in the Lorenz and 2D Navier–Stokes equations. *Physica D: Nonlinear Phenomena*, 240(18):1416–1425.

[48] Helin, T. and M. Burger (2015). Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems. *Inverse Problems*, 31(8):085009.

[49] Houtekamer, P. and J. Derome (1995). Methods for ensemble prediction. *Monthly Weather Review*, 123(7):2181–2196.

[50] Houtekamer, P. and H. Mitchell (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, 126(3):796–811.

[51] Iglesias, M., K.J.H. Law, and A.M. Stuart (2013). Ensemble Kalman methods for inverse problems. In *Inverse Problems*, 29(4): 134–141.

[52] Iglesias, M., K. Lin, and A.M. Stuart (2014). Well-posed Bayesian geometric inverse problems arising in subsurface flow. In *Inverse Problems*, 30(11): 114001.

[53] Iglesias, M. (2015). Iterative regularization for ensemble data assimilation in reservoir models. In *Computational Geosciences*, 19(1): 177–212.

[54] Iglesias, M. (2016). A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems. In *Inverse Problems*, 32(2).

[55] Jazwinski, A (2007). Stochastic processes and filtering theory. *Courier Corporation*.

[56] Johansen, A. and A. Doucet (2008). A note on auxiliary particle filters. *Statistics & Probability Letters*, 78(12):1498–1504, 2008.

[57] Kalman, R (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.

[58] Kalman, R. and R. Bucy (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83(1):95–108.

[59] Kalnay, E (2003). Atmospheric modeling, data assimilation and predictability. *Cambridge University Press*.

[60] Kantas, N., A. Beskos, and A. Jasra (2014). Sequential Monte Carlo methods for high-dimensional inverse problems: A case study for the Navier Stokes equations. *SIAM Journal on Uncertainty Quantification*, 2(1):464–489.

[61] Kawai, R. (2017). Adaptive importance sampling Monte Carlo simulation for general multivariate probability laws. *Journal of Computational and Applied Mathematics*, 319:440–459.

[62] Kaipio, J. and E. Somersalo (2006). Statistical and Computational Inverse Problems, volume 160. *Springer Science & Business Media*.

[63] Kelly, D., K. Law, and A.M. Stuart (2014). Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time. *Nonlinearity*, 27(10):2579.

[64] Kelly, D. and A. Stuart (2016). Ergodicity and accuracy of optimal particle filters for Bayesian data assimilation. *arXiv preprint arXiv:1611.08761*.

[65] Knapik, B., A. van der Vaart, J. van Zanten (2011). Bayesian inverse problems with Gaussian priors. *The Annals of Statistics*, 39(5):2626–2657.

[66] Giles, M. (2015). Multilevel monte carlo methods. *Acta Numerica*, 24:259–328.

[67] Gland, F., V. Monbet, and V. Tran (2009). Large sample asymptotics for the ensemble Kalman filter. *PhD thesis*, INRIA.

[68] Law, K.J.H., D. Alonso, A. Shukla, and A.M. Stuart (2016). Filter accuracy for the Lorenz 96 model: Fixed versus adaptive observation operators. *Physica D: Nonlinear Phenomena*, 325:1–13.

[69] Law, K.J.H., A. Shukla, and A.M. Stuart (2012). Analysis of the 3DVAR filter for the partially observed Lorenz'63 model. *Discrete and Continuous Dynamical Systems A*, 34(2014), 1061-1078.

[70] Law, K.J.H., A.M. Stuart and K. Zygalakis (2015). Data Assimilation: A Mathematical Introduction. *Springer*.

[71] Lindvall, T. (2002). Lectures on the Coupling Method. *Springer*.

[72] Lorenc, A. (1986). Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474):1177–1194.

[73] Lieberman, C. K. Willcox and O. Ghattas (2010). Parameter and state model reduction for large-scale statistical inverse problems. *SIAM Journal on Scientific Computing*, 32(5):2523–2542.

[74] Lindvall, T. (2002). Lectures on the coupling method, *Courier Corporation*.

[75] Lorentzen, R., K. Fjelde, J. FrØyen, A. Lage, G. Naevdal and E. Vefring (2001). Underbalanced and low-head drilling operations: Real time interpretation of measured data and operational support, *SPE Annual Technical Conference and Exhibition*.

[76] Lu, Y., A.M. Stuart and H. Weber (2016). Gaussian approximations for probability measures on $\mathbb{R}^d$. *SIAM Journal on Uncertainty Quantification*, 5: 1136–1165.

[77] MacKay D (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

[78] Majda A. and J. Harlim (2012). *Filtering complex turbulent systems*. Cambridge University Press.

[79] Martin, J., L. Wilcox, C. Burstedde and G. Omar (2012). A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487.

[80] Marzouk, Y. and D. Xiu (2009). A stochastic collocation approach to Bayesian inference in inverse problems. *Communications in Computational Physics*, 6(4):826–847.

[81] Mattingly, J., A.M. Stuart, and D. Higham (2002). Ergodicity for PDE's and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and Their Applications*, 101(2):185–232.

[82] Meyn, S. and R. Tweedie (2012). Markov chains and stochastic stability. *Springer Science & Business Media*.

[83] Moodey, A., A. Lawless, R. Potthast, and P. Van Leeuwen (2013). Nonlinear error dynamics for cycled data assimilation methods. *Inverse Problems*, 29(2):025002.

[84] Naevdal, G., T. Mannseth, and E. Vefring and V. Garcia (2002). Near-Well Reservoir Monitoring Through Ensemble Kalman Filter - EnKF. *Proceeding of SPE Improved Oil Recovery Symposium*.

[85] Nickl, R. (2017). Bernstein-von mises theorems for statistical inverse problems : Schrö odinger equation. *arXiv preprint arXiv:1707.01764*.

[86] Nielsen, F. and V. Garcia (2009). Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*.

[87] Oliver, D., A. Reynolds, and N. Liu (2008). Inverse theory for petroleum reservoir characterization and history matching. *Cambridge University Press*.

[88] Petersen, K. and M. Pedersen (2008). The matrix cookbook. *Technical University of Denmark*, 7:15.

[89] Pinski, F., G. Simpson, A.M. Stuart, and H. Weber (2015). Kullback–Leibler approximation for probability measures on infinite dimensional spaces. *SIAM Journal on Mathematical Analysis*, 47(6):4091–4122.

[90] Pinski, F., G. Simpson, A.M. Stuart, and H. Weber (2015). Algorithms for Kullback–Leibler approximation of probability measures in infinite dimensions. *SIAM Journal on Scientific Computing*, 37(6):A2733–A2757.

[91] Pitt, M. and N. Shephard (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599.

[92] Rauch, H., C. Striebel, and F. Tung (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450.

[93] Rebeschini, P and R. Van Handel (2015). Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, 25(5):2809–2866.

[94] Reich, S (2011). A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, 51(1):235–249.

[95] Reich, S. and C. Cotter (2015). Probabilistic forecasting and Bayesian data assimilation. *Cambridge University Press*.

[96] Sanz-Alonso, D. and A. Stuart (2015). Long-time asymptotics of the filtering distribution for partially observed chaotic dynamical systems. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1): 1200–1220.

[97] Sanz-Alonso, D. and A. Stuart (2016). Gaussian Approximations of Small Noise Diffusions in Kullback-Leibler Divergence. *In Communications in Mathematical Sciences*, 15(7).

[98] Schillings, C. and A.M. Stuart (2017). Analysis of the ensemble Kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, 55(3):1264–1290.

[99] Skjervheim, J., G. Evensen, S. Aanonsen, B. Ruud, and T. Johansen (2011). Incorporating 4D seismic data in reservoir simulation models using ensemble Kalman filter. In *SPE Journal*, 12(3): 282–292.

[100] Snyder C., T Bengtsson, P. Bickel, and J. Anderson (2008). Obstacles to high-dimensional particle filtering. In *Monthly Weather Review* 136(12):4629–4640.

[101] Snyder C. (2011). Particle filters, the "optimal" proposal and high-dimensional systems. In *Proceedings of the ECMWF Seminar on Data Assimilation for Atmosphere and Ocean*, 1–10.

[102] Stuart, A.M. (2010). Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559.

[103] Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. SIAM.

[104] Tokdar, S. and R. Kass (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60.

[105] Tong, X., A. Majda, and D. Kelly (2015). Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation. *Nonlinearity*, 29(2).

[106] Tong, X., A. Majda, and D. Kelly (2016). Nonlinear stability and ergodicity of ensemble based Kalman filters. *Nonlinearity*, 29(2):657.

[107] Van der Vaart, A. (1998). Asymptotic statistics, volume 3. *Cambridge University Press*.

[108] Van Leeuwen, P. (2010). Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, 136(653):1991–1999.

[109] Van Leeuwen, P., Y. Cheng and S. Reich (2015). Nonlinear data assimilation. *Springer*.

[110] Van Leeuwen, P. (2010). Towards adjoint-based inversion for rheological parameters in nonlinear viscous mantle flow. *Physics of the Earth and Planetary Interiors*, 234:23–34.